Trabajo de Investigación de Maestría:

Detección del carcinoma de glándula mamaria fusionando variables clínicas y termográficas

PRESENTADO ANTE:

Comité Asesor de Posgrado E³T

Por:

Edwin Santiago Alférez Baquero

Universidad Industrial de Santander

ESCUELA DE INGENIERÍAS
ELÉCTRICA, ELECTRÓNICA
Y DE TELECOMUNICACIONES

Bucaramanga, marzo de 2010

DETECCIÓN DEL CARCINOMA DE GLÁNDULA MAMARIA FUSIONANDO VARIABLES CLÍNICAS Y TERMOGRÁFICAS

EDWIN SANTIAGO ALFÉREZ BAQUERO

Trabajo de investigación desarrollado para optar al título de Magister en Ingeniería Electrónica

Director:
PhD. OSCAR GUALDRÓN GONZÁLEZ

Codirectora:
MD. OLGA MERCEDEZ ÁLVAREZ OJEDA

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERÍAS FISICO MECÁNICAS
ESCUELA DE INGENIERÍAS ELÉCTRICA, ELECTRÓNICA Y TELECOMUNICACIONES
BUCARAMANGA
2010

AGRADECIMIENTOS

Agradezco a Dios por todas las metas alcanzadas hasta el momento, en especial de la culminación de este proyecto de grado. A mi director Oscar Gualdrón por guiarme a través del desarrollo del proyecto y por su esfuerzo para aconsejarme y corregir mis errores. A la doctora Olga Álvarez, codirectora, agradezco todo su apoyo logístico a través del proceso de captación de pacientes y a su participación en la planeación original del proyecto de Colciencias. Al doctor Álvaro Niño, quien fue el principal médico Oncólogo que aportó en la consecución de las pacientes que participaron en el estudio. Al doctor Insuasti por permitir realizar muchos de los registros en su consultorio. A los estudiantes de medicina: Jenny, Hipólito, Rubén, Viviana y Sebastián por su colaboración en el registro termográfico. Al doctor Luis Orozco por sus aportes y observaciones en el desarrollo de esta tesis. A Leandro Ariza por sus contribuciones y consejos al comenzar mis cursos de maestría. A Victor Martinez por brindarme su amistad y aportarme varías ideas en este proyecto. A los doctores Arturo Plata y Jaime Meneses por sus valiosa revisión y calificación de este trabajo. A todos mis amigos del grupo CPS por la calidez y el ánimo que me trasmitían. A Cristian y Olga Sofía, mis estudiantes de pregrado, por su contribución a la base de datos de las variables clínicas. A Raquel Maldonado por sus servicios de auxiliar para contactar y verificar a las pacientes con sospecha de cáncer de mama. Agradezco a mi novia Claudia Durán por toda la paciencia y el apoyo mientras trabajaba en este proyecto. A doña Flor Ardila quien siempre me tendió la mano cuando más lo necesité. A mi Madre quien me formó como persona y me ayudo a ser un profesional. Y a todas las personas que de una u otra forma aportaron en el desarrollo de este trabajo, pero que olvido mencionarlas.

Para las dos mujeres de mi vida:
Mi madre y mi kiki

Tabla de contenido

Lista de figuras

Lista de Tablas

Lista de Anexos

RESUMEN

TÍTULO: DETECCION DEL CARCINOMA DE GLÁNDULA MAMARIA FUSIONANDO VARIABLES CLÍNICAS Y TERMOGRÁFICAS[*]

AUTOR: Edwin Santiago Alférez Baquero[†]

PALABRAS CLAVES: Termografía infrarroja, Biodatos, Cáncer de Seno, Procesamiento digital de imágenes, Curva ROC, Transformada de Hough, Red neuronal artificial (RNA), Fuzzy c-means.

DESCRIPCIÓN
Este trabajo busca clasificar una muestra poblacional de mujeres, en pacientes con y sin cáncer de mama, teniendo en cuenta la información termográfica de los senos y los datos clínicos de cada paciente. En primer lugar, se registraron los termogramas de mama y se recopiló la información clínica de cada paciente. Posteriormente, se eligió de forma manual y automática las regiones en la termografía que contienen las glándulas mamarías. En la segunda técnica se utilizó principalmente la transformada parabólica de Hough. A partir de estas zonas, se extrajeron descriptores de primer y segundo orden. Estos últimos se calcularon a partir de la matriz de co-ocurrencia, en 4 direcciones diferentes: horizontal, diagonal a 45º, vertical y diagonal a 135º. Asimismo, se preseleccionan los datos clínicos que tienen relevancia en el estudio. A continuación, se procedió a realizar la selección de variables que entran al sistema de clasificación. Se planearon tres técnicas: correlación lineal, selección secuencial de variables y discriminación mediante el área bajo la curva ROC. Sin embargo, este último es el que se aplica antes de la clasificación. Posteriormente, se aplicaron cuatro diferentes algoritmos de clasificación para cada grupo de datos: una RNA backpropagation, una red neuronal con correlación en cascada, una RNA probabilística y la técnica de agrupamiento Fuzzy C-means. Por último, se determinó la sensibilidad, especificidad y área bajo la curva ROC de cada prueba de clasificación y se compararon los resultados para cada tipo de algoritmo implementado.

[*]Proyecto de grado
[†]Facultad de ingenierías físico mecánicas, Escuela de Ingenierías Eléctrica, Electrónica y Telecomunicaciones, Director: Oscar Gualdrón González, Codirectora: MD. Olga M. Álvarez Ojeda.

ABSTRACT

TITLE: BREAST CANCER DETECTION COMBINING THERMOGRAPHICAL AND CLINICAL VARIABLES[*]

AUTOR: Edwin Santiago Alférez Baquero[†]

KEYWORDS: Infrared thermography, Biodata, Breast Cancer, Digital Image Processing, ROC curve, Hough transform, Artificial Neural Network (ANN), Fuzzy c-means.

This study aims to classify a sample of patients with and without breast cancer, taking into account the thermographical information of their breasts and their clinical data as well. At first, the breast thermograms and clinical information were collected from each patient. After this, regions were selected manually and then automatically, in order to enclose the breast-area, mainly using the parabolic Hough transform. From these areas, first and second order thermal descriptors were extracted. The second order descriptors were calculated from the co-occurrence matrix in 4 different directions: horizontal, diagonal (at 45 degrees), vertical and diagonal (at 135 degrees). Also, some clinical data relevant for this study was pre-selected. After this, the selection of variables for the classification system was made using three techniques: linear correlation, sequential selection of variables and discrimination, applying the area under the ROC curve method (the one applied before classification). Later on, four different classification algorithms for each data set were used: an ANN back propagation, a neural network with cascade correlation, a probabilistic neural network and Fuzzy C-means clustering. Finally, the sensitivity, specificity and area under the ROC curve were determined for each classification test and the results were compared for each type of algorithm implemented.

[*]Grade project

[†]Mechanical physics engineering faculty, School of Electrical Engineering, Electronics and Telecommunications, Director: Oscar Gualdrón Gonzalez, Co-Director: MD. Olga M. Álvarez Ojeda.

INTRODUCCIÓN

El cáncer de seno es un problema de salud mundial, que afecta a las mujeres sin distinción alguna de raza o sociedad. En países desarrollados la atención en salud proporciona un rápido cuidado y un seguimiento constante de las pacientes. Aún así, sigue siendo uno de los cánceres más frecuentes [1]. En Colombia, la población tiene acceso limitado a exámenes, debido al propio sistema de salud, a la disponibilidad del mismo sólo en ciertas zonas geográficas y a la cantidad de mujeres sin cobertura por parte de una entidad promotora de salud EPS; en muchos casos, cuando una mujer del campo o de bajos recursos, llega por un dolor o alguna secreción al médico especialista, es demasiado tarde como para realizar un tratamiento sobre esta enfermedad, que en este punto es terminal.

La termografía infrarroja es una técnica no invasiva, de bajo costo operativo, que se caracteriza por encontrar anomalías en el patrón térmico mucho antes de desarrollarse completamente el carcinoma. Por lo tanto, la termografía infrarroja constituye un examen fisiológico, alternativo, de alto poder predictivo, que podría volverse una herramienta de primera línea en la detección del cáncer de seno. Los procedimientos tradicionales del examen de termografía son realizados a través de la percepción del médico por medio de algún software que describa la radiación mostrando las temperaturas. Estos utilizan criterios cualitativos de simetría o variaciones termográficas cuantitativas, que en cualquier caso dependen de la concepción y de la experiencia del médico. El subjetivismo presente en el examen termográfico usual, limita su aplicación y su efectividad. Por otro lado, el diagnóstico general del estado normal o anormal de la enfermad, también depende de factores no termográficos, como la historia familiar, el tamaño del seno, el reemplazo hormonal, entre otros. Surge entonces la pregunta: *¿La combinación de variables termográficas y clínicas a través de un sistema basado en inteligencia artificial, aumentará la capacidad de detección del carcinoma de glándula mamaria?*

La solución a esta pregunta se aborda a lo largo de este trabajo. Este documento se encuentra divido en 8 capítulos. En el primer capítulo se describen los fundamentos teóricos que se emplearan a lo largo del trabajo, iniciando por el concepto de la termografía infrarroja, pasando por las nociones médicas referentes a la anatomía y patología de la glándula mamaria, hasta las bases de procesamiento digital de imágenes e inteligencia artificial para la clasificación y reconocimiento. El segundo capítulo relata el proceso de adquisición de la información, la cual está compuesta de dos partes fundamentales: el registro de termografía infrarroja y los formularios que conllevan la información sociodemográfica, hereditaria, hormonal y fisiológica. En este apartado, se explican el protocolo de adquisición de las termografías y la cantidad de variables clínicas recopiladas. El tercer capítulo, detalla el algoritmo de segmentación de las glándulas mamarias. Este se compone de una fusión de contornos del torso y la geometría parabólica determinada mediante la transformada de Hough. La selección de las variables es descrita en el capítulo 4, implementando tres técnicas: los coeficientes de correlación lineal Pearson y Spearman, el área bajo la curva ROC y la selección secuencial de variables. En el capítulo 5 se relata como se divide la información en diferentes conjuntos de datos, dependiendo de la dirección de la matriz de co-ocurrencia de los descriptores de segundo orden. Estos son utilizados para entrenar y probar cada algoritmo de clasificación, como señala el capítulo 6. Se prueban cuatro metodologías, tres redes neuronales artificiales: la arquitectura feedfordward Backpropagation, la correlación en cascada, redes probabilísticas y el agrupamiento con fuzzy c-means. A lo largo de esta sección, se determinan los parámetros de sensibilidad, especificidad y área bajo la curva ROC, para cada prueba y algoritmo implementado. De esta forma se validan los sistemas de clasificación. En el capítulo 7 se presentan las conclusiones de toda la investigación y en el 8 se referencia toda la literatura que ayudo a construir el proyecto.

1 MARCO TEÓRICO

1.1 TERMOGRAFÍA INFRARROJA

La termografía infrarroja es una técnica que permite identificar, a distancia y sin contacto físico con el objeto o cuerpo, cualquier anomalía que se manifieste como un cambio de la temperatura, midiendo los niveles de radiación infrarroja.

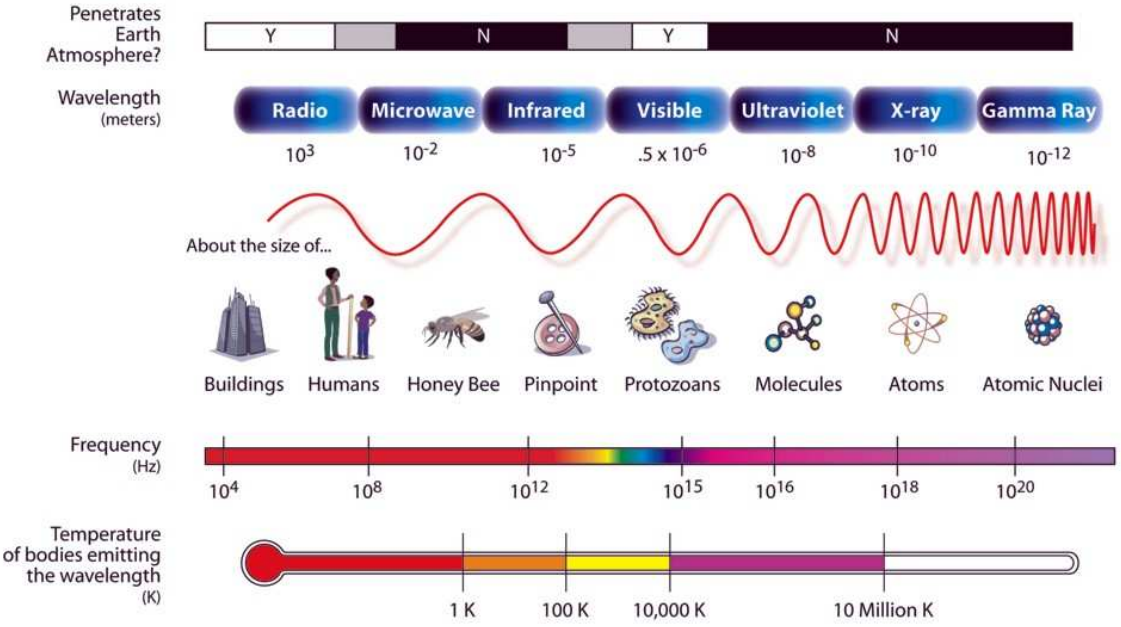
En lo que respecta a esta técnica, la inspección de los cuerpos u objetos se realiza generalmente mediante una cámara termográfica, que cuenta con un sistema de adecuación de señales que plasman la radiación percibida, en un arreglo matricial de sensores, como datos de temperatura (teniendo en cuenta la longitud de onda de onda emitida por el cuerpo para medir la temperatura). Para facilitar su análisis, estas matrices pueden emular una fotografía, donde cada dato de temperatura es manejado como un nivel de intensidad lumínica en un mapa de colores [1].

Como se observa en Figura 1, la radiación infrarroja ocupa el rango desde los 300GHz (1mm) hasta los 400THz (750nm) aproximadamente. Todo cuerpo emite este tipo de radiación, por lo cual es muy utilizada en la industria, astronomía, medicina, entre otras. Este segmento del espectro puede ser dividido en cuatro regiones: onda corta (1.1 – 2.5μm), onda media (2.5 – 7.5μm), onda larga (7 - 15μm) y onda muy larga (> 15μm).

1.1.1 Termografía infrarroja en aplicaciones médicas

La termografía infrarroja es una técnica que registra la temperatura de la superficie de un cuerpo y puede ofrecer información sobre las disfunciones térmicas asociadas con algún tipo de patología. El registro de la temperatura corporal puede ofrecer una información invaluable de los procesos fisiológicos que

Figura 1. Espectro Electromagnético



Fuente: http://mynasadata.larc.nasa.gov/ElectroMag.html

pueden presentar anormalidad. La disipación de calor a través de la piel ocurre a través de la radiación de energía del espectro, lo cual hace a los detectores infrarrojos de gran utilidad en el proceso de localización de anormalidades patológicas.

Los termogramas clínicos infrarrojos son difíciles de identificar debido a que las variaciones de temperatura en el cuerpo humano pueden ser muy ligeras y se pueden presentar alteraciones en la temperatura corporal por factores externos, como la temperatura ambiente. Gracias al rápido avance en las tecnologías para adquisición y procesado de imágenes infrarrojas, en la actualidad existen muchas aplicaciones en el campo de la medicina [3], entre las cuales se puede nombrar:

- Oncología (estudio de las neoplasias malignas y benignas, basado en quimioterapia, radioterapia y cirugía oncológica)
- Dolencias

- Desordenes vasculares (alteración de los vasos sanguíneos)
- Artritis/Reumatismo (trastornos articulares)
- Neurología (estudios y tratamientos sobre el sistema nervioso)
- Cirugía
- Oftálmico (estudios y tratamientos en los ojos)
- Viabilidad tisular (relacionado con el estudio de el funcionamiento y condición de los tejidos)
- Desordenes dermatológicos (alteraciones en los tejidos de la piel)
- Monitoreo de la eficacia de medicinas y tratamientos
- Tiroides (estudio de la glándula que produce hormonas)
- Odontología (estudio y tratamiento del aparato estomatognático)
- Respiratorio
- Medicina deportiva y de rehabilitación

1.1.1.1 Ventajas de la termografía infrarroja

La termografía infrarroja detecta cambios locales en la fisiología, el metabolismo o el flujo de la sangre antes de que otro examen clínico o prueba de chequeo pueda detectarlos [4]. La termografía y la mamografía son técnicas que no diagnostican específicamente el cáncer de seno, pero son capaces de detectar cambios en los tejidos por diferentes causas. La mamografía detecta cambios anatómicos y aunque es más precisa en la identificación exacta de la localización de una lesión, la termografía detecta cambios fisiológicos mucho antes de que ocurran los cambios anatómicos. Se ha demostrado que la termografía puede detectar estados precancerosos del tejido del seno, mucho tiempo antes de que se puedan identificar por otros método [5], [6]. La termografía posee la habilidad de detectar muchos de los tumores más pequeños que difícilmente podrían ser palpados mediante el examen clínico de los senos [7].

1.1.1.2 Termografía de seno

La termografía es una técnica utilizada para la valoración de mujeres de forma complementaría a la mamografía o la ecografía. Si se realiza un chequeo termográfico unido a la valoración médica y a la mamografía se puede proporcionar un alto valor de confiabilidad de que el cáncer de seno pueda ser detectado.

A través de los años, numerosos estudios han demostrado las virtudes de la termografía infrarroja en el ámbito de la medicina como una técnica de diagnóstico alternativa, no invasiva e inocua, basada en la transferencia de calor por medio de radiación térmica. Asimismo, se han realizado estudios que han puesto en duda la capacidad de esta técnica de llegar a ser un sustituto adecuado de la mamografía, como el desarrollado por el Instituto Nacional de Cáncer de Estados Unidos desde 1973 a 1981 conocido por sus siglas en inglés como BCDDP[5] [8].

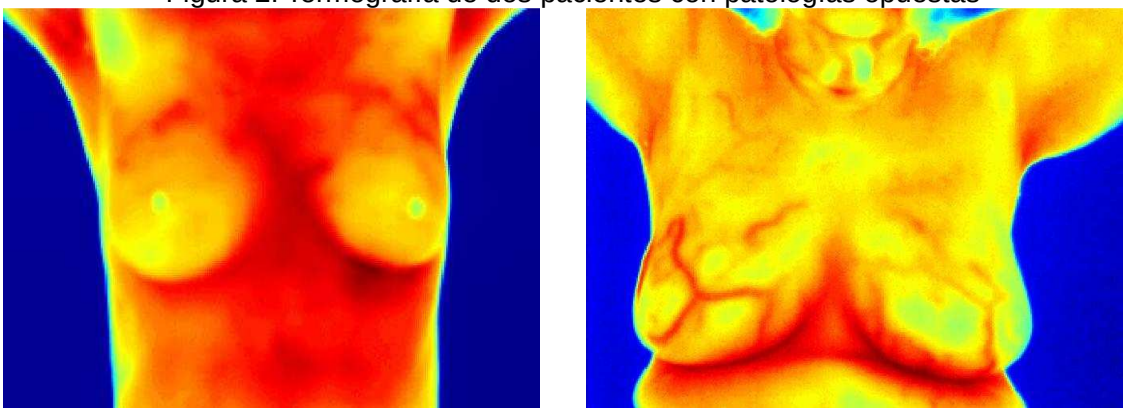
1.1.1.3 Detección de cáncer de seno a través del análisis de simetría térmica

Una condición normal detectada a través de termografía infrarroja, encuentra una buena simetría térmica entre los senos, sin una elevación anormal de la temperatura en una zona en especial (Figura 2a). Una condición de un cáncer infamatorio en una etapa avanzada de la enfermedad, se puede observar debido a la aparición de anormalidades térmicas en una región específica del seno y en una elevación de la temperatura comparada a la del seno contrario (Figura 2b) [9] [9].

Las aplicaciones en termografía infrarroja, deben estar orientadas a ser procesos objetivos y automáticos que eliminan la repercusión e influencia del factor humano en los resultados obtenidos. Estas implementaciones llevan a cabo los siguientes procedimientos básicos [11], [12], [13]:

[5] En inglés, Breast Cancer Detection and Demonstration Project

Figura 2. Termografía de dos pacientes con patologías opuestas



a. Paciente sana b. Paciente con cáncer

Fuente: el autor

1. Segmentación de las regiones correspondientes a los senos en el termograma.
2. Mediciones estadísticas (descriptores) por separado sobre el conjunto de píxeles que conforman cada región segmentadas (mamas), las cuales se comparan entre si con el fin de identificar una posible asimetría térmica.
3. Con los descriptores estadísticos extraídos se procede a realizar un algoritmo de clasificación (redes neuronales, lógica fuzzy, entre otros) para evaluar la termografía infrarroja como técnica diagnóstica.

1.1.1.4 Validez de la termografía infrarroja en la detección del carcinoma de glándula mamaria.

Son diversos los resultados encontrados en cuanto a la eficiencia de esta técnica para la detección del cáncer de seno. Parisky [14], reporta a través de su procedimiento, un 97% de sensibilidad[6], un 14% de especificidad[7], un 95% de valor predictivo negativo y un 24% de valor predictivo positivo, que muestran que la termografía infrarroja es efectiva en el diagnóstico del cáncer de seno.

[6] Probabilidad de clasificar correctamente a una paciente enferma

[7] Probabilidad de clasificar correctamente a una paciente sana

Keyserlingk [15] reportó un 83% de sensibilidad del examen termográfico, combinando los mamogramas sospechosos con los termogramas anormales, la sensibilidad se incrementó al 93% y al 98% cuando los exámenes clínicos sospechosos también fueron tenidos en cuenta. Head [16], demuestra que mujeres con imágenes infrarrojas anormales tienen un incremento en el riesgo, aproximadamente del 30%, de desarrollar cáncer de seno. Señalaron que la técnica de termografía tiene un alto valor predictivo. De esta forma, probablemente pueda ser integrado con otros indicadores pronósticos para decidir si se realiza tratamiento de quimioterapia.

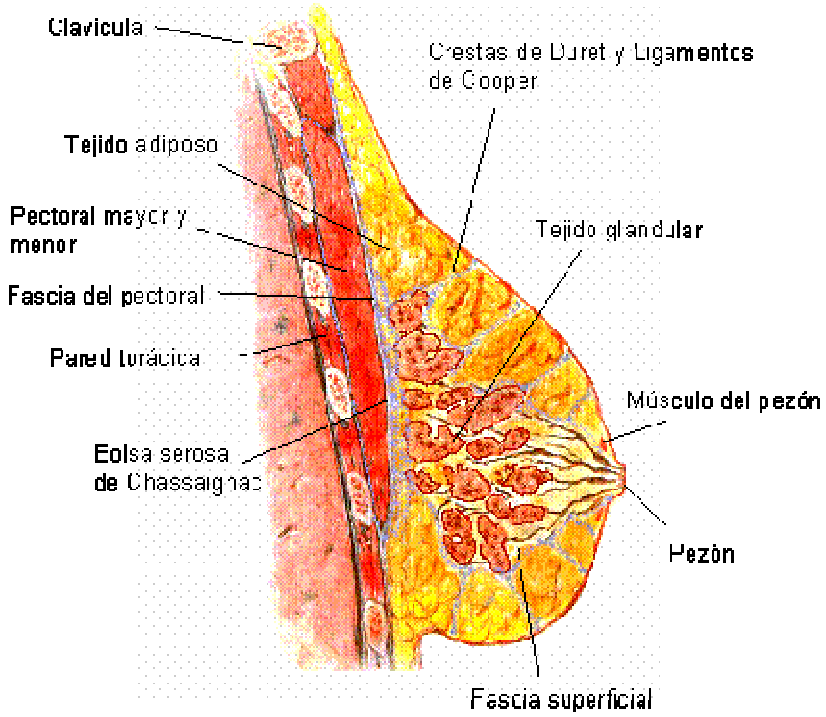
1.2 ANATOMÍA Y PATOLOGÍA DE LA GLÁNDULA MAMARIA

Este trabajo de investigación, busca determinar las características termográficas y clínicas que permitan la mejor descripción a la hora de establecer una condición de anormalidad en las glándulas mamarias, especialmente en el posible diagnóstico de neoplasias malignas. Por ende, se abordarán algunos conceptos fundamentales sobre la fisiología, patología y evidencia clínica de la glándula mamaria.

1.2.1 Estructura de la glándula mamaria

La glándula mamaria es un conjunto especializado de glándulas que durante el desarrollo femenino modifican su secreción con el fin de producir leche. Se le da el nombre de mama, a la región antero superior lateral del tronco humano [17], ubicada entre la segunda o tercera costilla extendiéndose hasta la sexta o séptima costilla. Por delante de la musculatura torácica y de la membrana conjuntiva que cubre los músculos, fijándolos a los huesos en la zona pectoral. Su extensión se da en sentido transversal, partiendo en el borde del esternón hasta la línea axilar anterior, con una prolongación axilar que se inicia en el cuadrante superior externo [18].

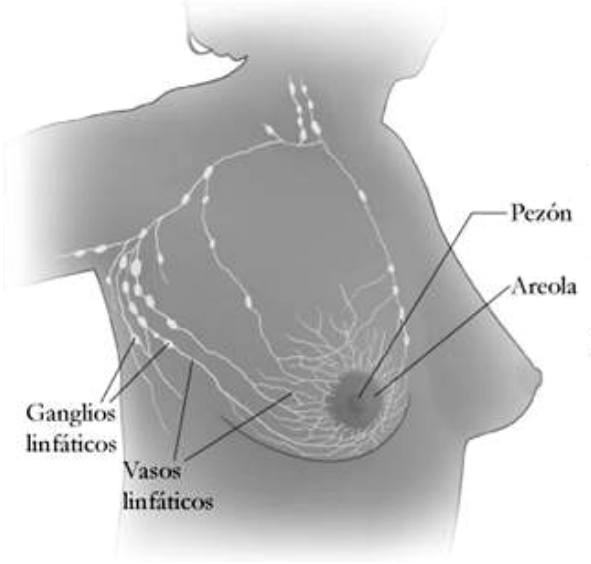
Figura 3. Corte transversal de la glándula mamaría



Fuente: [22]

La Figura 3 muestra la glándula mamaria, cuyo aspecto exterior es una protuberancia de tamaño y abultamiento variable, consta de estructuras de tipo externas e internas [19]. La glándula mamaria está constituida básicamente por dos elementos: los acinos glandulares y los ductos. El primero contiene las células productoras de leche, y el segundo es un arreglo de estructuras tubulares y huecas, que se ramifican en forma de árbol, confluyendo progresivamente en canalículos cada vez más gruesos, terminando en uno de los doce a dieciocho vértices galactóforos [20]. Los galactóforos son dilataciones ductales a modo de receptáculos que se encuentran ubicados inmediatamente por detrás del pezón, formados por un epitelio escamoso sin queratina [21]. Entre las estructuras externas de la glándula mamaria la más importante, es la región pezón-areola. En ésta se encuentran un tipo de células llamadas mioepiteliales, que tienen como

Figura 4. Ganglios y vasos linfáticos alrededor de la mama



Fuente: http://www.meb.uni-bonn.de/cancer.gov

característica especial la elongación, es decir, la capacidad de contraerse y estirarse a manera de fibras musculares. Debido a la cantidad de terminales nerviosas conectadas y a la presencia de fibras musculares lisas y con forma radial, en la región pezón-areola, se presentan erecciones debido a un tacto suave o succión. Entre los tejidos que componen el resto de las mamas están: el tejido conjuntivo[8], el tejido adiposo[9] y la aponeurosis[10] conocida como ligamento de Cooper. La proporción relacional entre la glándula y tejido graso es de uno a uno en mujeres que no están lactando, y de dos a uno en mujeres en estado lactante [19].

Aproximadamente las tres cuartas partes de la linfa que sale de la zona mamaria se dirige a los ganglios linfáticos ubicados en la axila más cercana a

[8]Colágeno y elastina

[9]Grasa

[10]Membrana conjuntiva que cubre los músculos y cuyas prolongaciones fijan éstos a los huesos

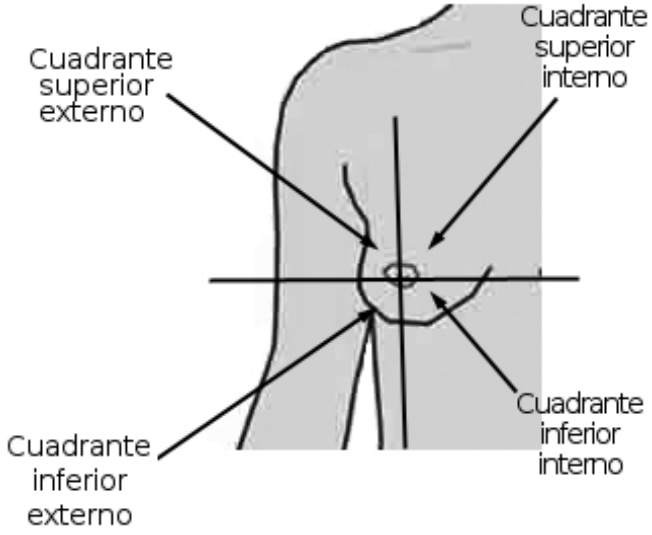
la mama. La otra parte es direccionada hacia los nódulos paraesternales, al seno opuesto y a los ganglios linfáticos en el abdomen. En los nódulos axilares se incluyen, el grupo transmuscular, subescapular que drena la parte interna de la glándula mamaria y el grupo humeral que drena el borde externo de la mama.

Como muestra la Figura 4, el drenaje linfático de las mamas drena en los ganglios linfáticos de la axila. Este drenaje tiene una especial relevancia en el ámbito oncológico, debido a que la región mamaria es un lugar de frecuente presencia de células cancerosas y si dichas células malignas se separan del tejido mamario, podrían fluir a otras partes del cuerpo vía el sistema linfático, produciendo lo que se conoce como metástasis.

1.2.2 División anatómica de la mama

La forma y tamaño de las mamas es diversa, condiciones propias como la edad, etnia, actividad atlética y hasta cirugías de tipo estético la hacen una zona morfológicamente variable. Su aspecto exterior no es un indicativo de gran peso de la anatomía interna o de su capacidad láctica. Asimismo, la forma de la glándula mamaria depende del tejido torácico sobre el cual se apoya y del soporte proporcionado por los ligamentos de Cooper. Cada mama se adhiere en su base a la pared torácica, por una envoltura del tejido conjuntivo profunda que recubre los músculos pectorales. En la parte superior de la zona pectoral se recibe soporte de la piel envolvente, esta combinación de soporte anatómico es lo que determina la forma de las mamas. Para determinar la simetría mamaria y la caída de la misma, la medida antropométrica de uso general es la estimación de la longitud de separación entre el pezón y el esternón Esta medida es relativa conforme a la edad, presentando en mujeres jóvenes un valor promedio de 0.21 metros.

Figura 5. División anatómica de la Mama



Fuente: http://meded.ucsd.edu/clinicalmed/breast.htm

En la Figura 5 se puede observar la subdivisión del seno en cuatro cuadrantes. El cuadrante superior externo, parte desde el punto medio del pezón hasta la axila. El cuadrante superior interno, parte del punto medio del pezón hasta el encuentro con el esternón. Estos dos cuadrantes en conjunto con el arreglo pezón-areola es por donde mayor circulación arterial se da en la glándula mamaria, puesto que por esta área están ligadas arterias que nacen de la arteria axilar. Los otros dos cuadrantes parten desde el punto medio del pezón hasta la zona de curvatura del seno y su composición en la mayoría de los casos es tejido adiposo.

1.2.3 Región de incidencia del cáncer de mama

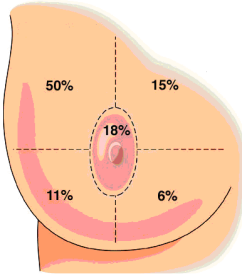
El proceso natural del cuerpo es renovar las células muertas, por células nuevas. Cuando este proceso se desordena, es decir, cuando se producen células nuevas y aún el cuerpo no las necesita, se empiezan a formar y acumular masas de tejido. Este tipo de masas se conoce como tumores, los

cuales pueden ser de tipo benigno o maligno. Dependiendo de la zona donde ocurra este fenómeno es la denominación que recibe. Los carcinomas de glándula mamaria se pueden clasificar en dos formas generales conforme el origen, los cercanos o sobre las células del epitelio ductal o los cercanos o sobre las células de los acino glandulares. En la Figura 6 se pueden observar los valores porcentuales de zonas de incidencia conforme el cáncer de seno. La zona superior exterior tiene una incidencia del 50%, por encontrarse cerca de los ganglios axilares. La segunda zona de mayor incidencia, se da en el conjunto pezón-areola por encontrarse los ductos y vasos linfáticos. La tercera zona es la superior interior con un 15 por ciento y las dos zonas inferiores tienen una incidencia menor por la poca influencia de ganglios y ductos.

1.2.4 Factores pronósticos

Un factor pronóstico es una característica que, por si sola o en combinación con otras, es capaz de dar información sobre la evolución clínica de un paciente. El uso de estos factores permite delimitar el grupo de pacientes con mayor riesgo de recaída, el subgrupo de más alto riesgo, y quiénes se beneficiarían de un tratamiento adyuvante oportuno y específico. En la tabla 1 se muestra una clasificación de algunos factores de riesgo para el cáncer de seno.

Figura 6. Incidencia del carcinoma de mama



Fuente: http://faculty.washington.edu/alexbert/MEDEX/

Tabla 1. Factores de riesgo de cáncer de seno.

Factor de riesgo		Característica
Sociodemográficos	Género	Ocurre unas 100 veces más en mujeres que en hombres [23].
	Edad	Las tasas de incidencia aumentan enormemente en edades entre los 45 y 50 años [24][25].
	Estado socioeconómico	Las mujeres de nivel socioeconómico alto tienen mayor riesgo de desarrollar cáncer de mama en comparación a las de estrato bajo [26].
	Área de residencia	Aumento de la incidencia en mujeres de áreas urbanas comparadas con las que residen en áreas rurales [27][28].
Hereditarios	Historia familiar de cáncer	Solo el 10% de mujeres diagnosticadas con esta lesión tienen antecedentes familiares positivos [29].
Hormonales	Edad de la menarquía	La menarquía temprana (antes de los 12 años) ha sido asociada a un incremento del riesgo en un 10 a 20% [30][31].
	Menopausia	Una menopausia tardía (después de los 54 años) incrementa el riesgo de desarrollar carcinoma mamario en un 3% por cada año que se tarde la menopausia [32][33].
	Embarazo temprano	Mujeres con un embarazo a término y paridad aumentada tienen un riesgo disminuido a la mitad del presentado en las nulíparas [25].
	Lactancia	Una lactancia prolongada ha demostrado ser un factor protector, disminuyendo el riesgo en un 3.4% por cada 12 meses de lactancia; adicionalmente, por cada parto el riesgo baja un 7.0% [34][35][36][37].

1.3 PROCESAMIENTO DIGITAL DE IMÁGENES (PDI)

Los detectores infrarrojos son capaces de medir la radiación térmica emitida por un objeto. Debido a la relación que existe entre la radiación infrarroja y la temperatura, se pueden medir y visualizar los diferentes perfiles térmicos a través de una imagen.

Figura 7. Etapas del procesamiento digital de imágenes

Fuente: el autor

Ésta es una matriz de datos, la cual puede contener el valor de las temperaturas registradas o los niveles de grises proporcionales a dichos valores. La mayoría de cámaras contienen software privativo, que restringe la información a un formato codificado o en su defecto, a un formato de imagen usual (.jpg, .tiff, .png,…).

La primera etapa para un procesamiento digital de termogramas, es el registro o adquisición de la imagen térmica, que se realiza con el detector o sensor de imágenes infrarrojas del sistema; este elemento es comúnmente una cámara infrarroja. La señal que recibe este sensor debe ser digitalizada a la salida del mismo. Después de que la imagen digital ha sido obtenida, el siguiente paso es el preprocesamiento de la imagen. La función de esta etapa es la de mejorar las características de la imagen térmica de manera que sirva para trabajar con éxito en los siguientes procesos. En el preprocesamiento típicamente se trabaja con técnicas para realzar el contraste, remover el ruido y características no deseadas en la imagen, entre otras operaciones requeridas.

El siguiente paso en el sistema de procesamiento es la segmentación. Definida en forma amplia, la segmentación realiza una partición del termograma de entrada en sus partes constituyentes u objetos. Posteriormente, la etapa de descripción, llamada también generación de características, trata con la extracción de los rasgos que resulta en alguna información cuantitativa de interés o características que son básicas para diferenciar una clase de objetos con otra (por ejemplo diferenciar los defectos térmicos o las zonas de altas temperaturas). El último estado comprende al reconocimiento y la interpretación. El reconocimiento es el proceso que etiqueta, o asigna un nombre, a un objeto basándose en la

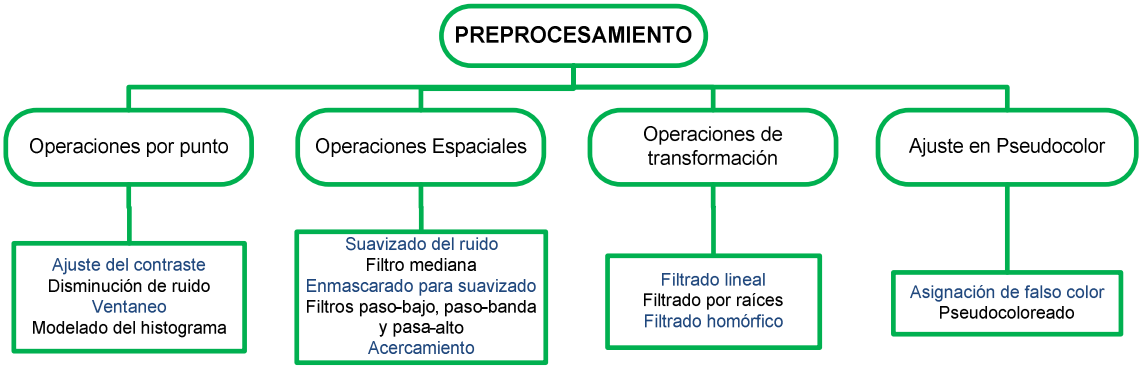
información que proveen sus descriptores. La interpretación involucra la asignación de significado a un conjunto de objetos reconocidos. Este proceso puede requerir técnicas de inteligencia artificial para la caracterización del problema (ver sección 1.4). La Figura 7 muestra un esquema que contiene las diferentes etapas del procesamiento de imágenes.

1.3.1 Preprocesamiento

El preprocesamiento atenúa o realza las características de una imagen, tales como bordes, límites o contraste. Este procedimiento no aumenta la información presente en los datos, pero sí incrementa el rango dinámico de las propiedades de una imagen para facilitar su manipulación. Entre los tipos de preprocesado se pueden encontrar: manejo del contraste y niveles de gris, disminución de ruido, suavizado, realces de bordes, filtrado, interpolación y ajustes de falso color [39],[40]. El principal problema al tratar de mejorar la imagen, es encontrar el ajuste adecuado. De esta forma, la mayoría de las técnicas empleadas son empíricas y se basan en procesos interactivos para obtener resultados coherentes. La Figura 8 muestra un esquema de las diferentes técnicas empleadas en el preprocesamiento.

En el tratamiento de imágenes termográficas se utiliza ajuste del contraste, suavizado del ruido, filtro gaussiano, asignación de falso color, entre otras. Sin embargo, dado que el termograma es una matriz de temperaturas, no es recomendable tratar estas imágenes directamente, pues modifican la información obtenida a partir de la radiación infrarroja. De esta forma, usualmente el preprocesamiento se efectúa sobre conversiones de las termografías a formatos estándares o a normalizaciones de la imagen (escala de grises). Con el fin, de mejorar la visualización o realizar posteriores procedimientos de segmentación.

Figura 8. Tipos de preprocesamiento



Fuente: adaptado de [38]

1.3.2 Segmentación

La segmentación es el proceso de dividir una imagen digital dentro de varios segmentos (regiones de pixeles). El objetivo de este procedimiento es simplificar y/o cambiar la representación de una imagen en "algo" que sea más fácil de analizar.

Generalmente, esta etapa se usa para localizar objetos o fronteras (líneas, curvas, etc.) dentro de la imagen. De forma más precisa, la segmentación en procesamiento digital de imágenes, es el proceso de asignar una etiqueta a cada pixel, tal que los puntos en una región sean similares, respecto a algunas características o propiedades como color, intensidad o textura. Además, las regiones adyacentes obtenidas, son ampliamente diferentes respecto a las mismas características [41].

La mayoría de los algoritmos de segmentación se fundamentan en dos propiedades básicas: la discontinuidad y la similaridad. De acuerdo a la primera, se divide la imagen basándose en los cambios abruptos de la intensidad, como por ejemplo los bordes. Con la segunda, se particiona la imagen en regiones que son similares, según un conjunto de criterios predefinidos. Ejemplos de esta clase

de procedimiento son: la umbralización, el crecimiento de regiones y, la división y fusión de regiones [40]. También, la combinación de estas dos clases de métodos puede mejorar sustancialmente la segmentación. Por ejemplo, es frecuente mezclar la umbralización junto a la detección de bordes.

1.3.3 Umbralización

La umbralización es la separación de objetos o regiones en una imagen con base en los niveles de grises, dependiendo de si están sobre o debajo de un valor determinado (umbral). Una simple definición matemática de la umbralización es:

$$g(x,y)=\begin{cases} G_0 \ si \ \ f(x,y)\geq T \\ G_B \ si \ \ f(x,y)<T \end{cases}$$

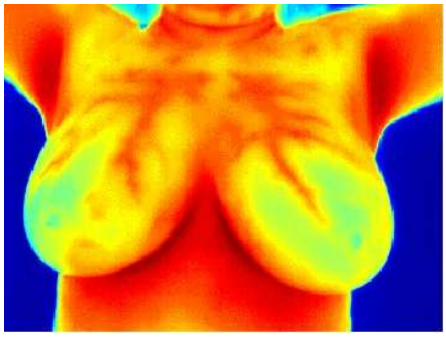
Donde G_0 es el valor de reemplazo para el objeto, G_B es el valor para sustituir al fondo y T es el valor de umbral. La salida de este procedimiento, es el 'objeto' o el 'fondo' etiquetado, el cual, debido a su naturaleza dicótoma, generalmente se representa como una variable Booleana '1' o '0'. En principio, la condición de prueba podría basarse sobre alguna otra propiedad además del brillo.

La pregunta central en la umbralización es: ¿Cómo elegir el umbral *T*? No existe un procedimiento universal para la selección de umbral que trabaje sobre todas las imágenes, pero hay una gran variedad de alternativas para automatizar el proceso.

1.3.3.1 Umbralización óptima

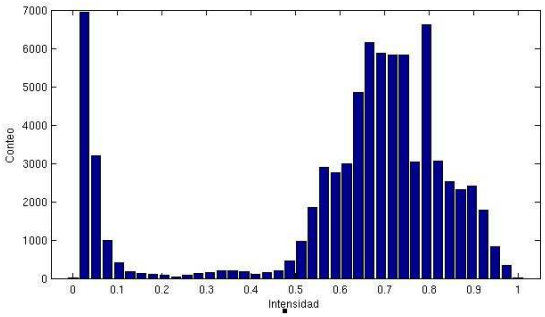
La umbralización óptima asume un histograma bimodal y señala al punto mínimo como el umbral. Por ejemplo, la Figura 9(a) muestra una termografía infrarroja de los senos de una mujer. Claramente se nota una diferencia entre el cuerpo de la persona y el fondo. El histograma puede revelar un gran detalle acerca de los

Figura 9. Ejemplo de umbralización óptima. Los niveles de intensidad (originalmente temperaturas) han sido normalizados entre 0 y 1.

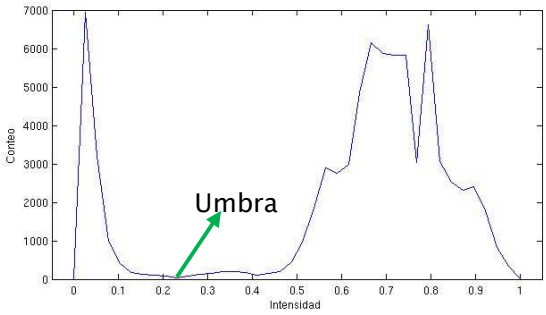


(a) Termografía de seno

(b) Binarización del termograma original



(c) Histograma de la termografía de seno



(d) Traza por segmentos del histograma

Fuente: el autor

objetos en una imagen, pero la interpretación del histograma no siempre es directa o útil. En la Figura 9(c) se ve que el histograma de la termografía es bimodal, puesto que existen dos picos bien distinguidos. Si se convierte el histograma a una función lineal por segmentos, entonces se puede observar el punto de transición entre las dos modas, como muestra la Figura 9(d). A este valor se le denomina “umbral óptimo”, que en este caso es de 0.23 (sobre intensidades normalizadas del termograma). El resultado de la segmentación se puede ver en la Figura 9(b). El algoritmo que busca el umbral óptimo, es aquél que determina el mínimo de una función, excluyendo los extremos. Este procedimiento asume que el histograma es bimodal y que los objetos de interés pertenecen a un lado del mínimo.

Figura 10. Segmentación de una termografía de seno con el método de Otsu

Fuente: el autor

1.3.3.2 Umbralización mediante varianza entre clases (método de Otsu)

La umbralización mediante varianza entre clases, fue propuesta por Otsu [42] y es llamada con frecuencia el "método de Otsu". El algoritmo maximiza la separación entre clases usando un criterio discriminante fundamentado en los momentos de cero y primer orden del histograma. El método no es trivial y se explica mejor en términos de la expresión usada para evaluarlo: la varianza entre clases. Esta es dada por la siguiente ecuación:

$$\sigma_B^{\,2}(k) = \frac{\left[\mu_T \omega(k) - \mu(k)\right]^2}{\omega(k)\left[1 - \omega(k)\right]} \qquad (1.1)$$

Los componentes de la varianza entre clases son:

Valor medio de la imagen → $\mu_T = \mu(L) = \sum_{i=1}^{L} i p_i$

Momento acumulativo de primer orden → $\mu(k) = \sum_{i=1}^{k} i p_i$

Momento acumulativo de orden cero → $\omega(k) = \sum_{i=1}^{k} p_i$

Donde *L* es el número de niveles de gris en la imagen y p_i es el número de pixeles en el nivel de gris *i*. Estas ecuaciones pueden parecer complicadas, pero una simple explicación de los momentos, es que estos suministran un valor escalar

que resume la relación entre los valores de pixeles y la posición espacial. Esto tiene sentido heurísticamente, cuando se recuerda que se esta intentando encontrar un umbral que produce la mejor separación entre objetos de la imagen.
En el método de Otsu, el umbral óptimo *k** es el nivel de gris cuando la varianza entre clases es máxima. Este ocurrirá (teóricamente) cuando el error cuadrado entre la media de la imagen, escalada por el momento de orden cero (una medida de dispersión del histograma) y el momento de primer orden (una medida de dispersión ponderada por el valor de gris), es grande. En otras palabras, un umbral que divide los valores de los pixeles en porciones ampliamente desiguales, evaluará el máximo de la varianza entre clases. La idea es que los objetos sean mínimamente dispersados en los niveles de gris, mientras el fondo será fuertemente dispersado. Se puede expresar el cálculo del umbral óptimo, *k**, como:

$$\sigma_B^{\,2}(k^*) = \max_{1 \le k \le L} \sigma_B^{\,2}(k) \qquad (1.2)$$

En la Figura 10 se muestra la termografía de la Figura 9(a) segmentada mediante el método de Otsu (con umbral de 0.39). Se nota claramente, que la umbralización logra una mejor descripción del cuerpo de la mujer.

1.3.4 Detección de Bordes mediante el algoritmo de Canny

La detección de bordes se refiere al proceso de identificar puntos, en una imagen digital, en los cuales el brillo de la imagen cambie bruscamente o presente discontinuidades. Existen muchas clases de algoritmos para la detección de bordes, pero la técnica más aplicada por su consistencia y efectividad, fue diseñada por Jhon F. Canny [43]. Este método, esta compuesto de varias etapas para lograr la detección de un amplio tipo de bordes.

1.3.4.1 Eliminación de ruido

El detector de bordes diseñado por Canny utiliza un filtro Gaussiano de primer

orden, debido a la usual presencia de ruido en las imágenes originales. De esta forma, la imagen se convoluciona con el filtro, resultando en una versión suavizada respecto a la original. Es importante considerar el tamaño del filtro gaussiano, puesto que afecta todo el proceso de detección de bordes (mediante la desviación estándar). Filtros pequeños causan menor suavizado y permiten la detección de líneas pequeñas y con gran curvatura. Filtros grandes causan un mayor suavizado, corriendo el valor de un pixel sobre una mayor área en la imagen. Este tipo de emborronamiento es útil en la detección de bordes grandes y gruesos.

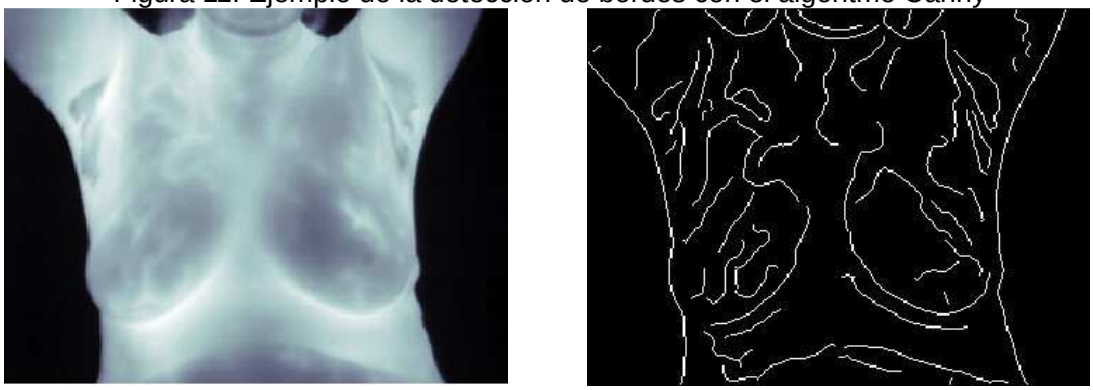
1.3.4.2 Determinación del gradiente

Si los niveles de intensidad de la imagen se consideran como una función bidimensional *f(x,y)*, el vector *gradiente* es definido por $\left[\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}\right]$. Donde sus componentes son derivadas parciales referentes a las direcciones horizontales y verticales. La dirección del máximo incremento de la función se expresa como $\tan^{-1}\left(\frac{\partial f}{\partial y} / \frac{\partial f}{\partial x}\right)$ y la magnitud del gradiente por $\sqrt{\left(\frac{\partial f}{\partial x}\right)^2 + \left(\frac{\partial f}{\partial y}\right)^2}$. El algoritmo de Canny usa diferentes aproximaciones del gradiente, para la detección de bordes en las direcciones: horizontal, vertical y diagonal (por ejemplo,mediante los operadores Roberts, Prewitt, Sobel, etc.) [39].

1.3.4.3 Supresión no máxima

Consiste en el adelgazamiento del ancho de los bordes, obtenidos con el gradiente, hasta lograr longitudes de un pixel. Para cada punto del borde, se observa un pixel en la dirección del gradiente y otro en la dirección inversa. Si la magnitud del pixel actual no es el máximo de los tres, entonces este punto no pertenece a un borde. Este método ayuda a localizar los verdaderos pixeles que pertenecen al borde.

Figura 11. Ejemplo de la detección de bordes con el algoritmo Canny



(a) Termograma de seno (b) Detección de bordes Canny

Fuente: el autor

1.3.4.4 Umbralización mediante histéresis

La imagen obtenida después de la supresión no máxima, puede contener muchos puntos que no corresponden a algún contorno. Para remover los puntos que pertenecen a falsos bordes, una apropiada umbralización es seleccionada. Así, todos los puntos que tengan una magnitud mayor al umbral, pueden ser conservados como bordes verdaderos, mientras otros son removidos como pixeles falsos. Entonces, si el umbral es pequeño, un número de puntos de bordes falsos pueden ser detectados como verdaderos. Sin embargo, algunos pixeles verdaderos pueden ser omitidos. Para evitar este problema, dos umbrales T_1 y T_2 pueden ser elegidos creando dos diferentes bordes de imágenes E_1 y E_2, donde $T_2 = 1.5T_1$. E_1 incluirá algunos puntos de bordes falsos, mientras que E_2 incluirá muy pocos. Se omitirán unos pocos bordes verdaderos. El algoritmo de selección de umbral comienza con puntos del borde en E_2, formando un contorno al conectar puntos de bordes adyacentes. El proceso continúa hasta que no haya más pixeles de bordes adyacentes. En la frontera del contorno, el algoritmo busca los próximos puntos de borde de la imagen E_1 y sus 8 más cercanos. Los espacios entre dos bordes pueden ser llenados tomando puntos de borde de E_1 hasta que el espacio haya sido completado. Este proceso perfecciona el contorno constituido por los bordes verdaderos de la Imagen.

La detección final se lleva a cabo comparando bordes del gradiente con el umbral. Este valor puede elegirse lo suficientemente pequeño, sólo cuando no haya ruido en la imagen. Así, todos los bordes verdaderos puedan ser descubiertos sin pérdida.

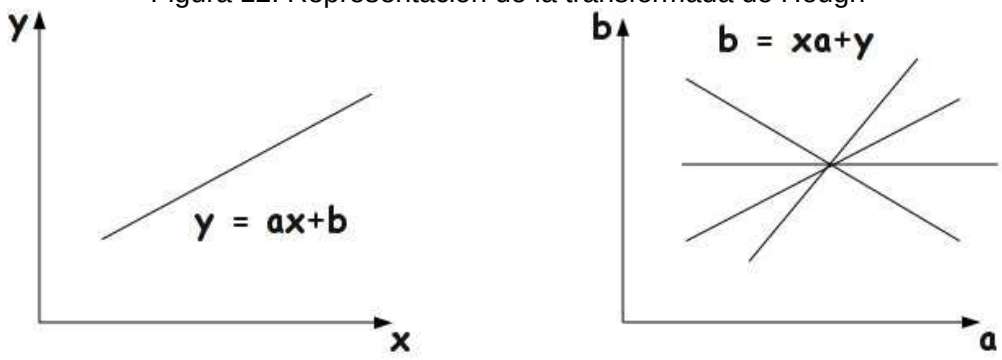
La Figura 11 muestra un ejemplo en la utilización del algoritmo desarrollado por Canny, en la detección de bordes de una termografía infrarroja de los senos de una mujer.

1.3.5 Transformada de Hough

Si los puntos de contorno encontrados son escasos, el resultado de la imagen de bordes puede consistir de puntos individuales en lugar de líneas rectas o curvas. De esta forma, para establecer una frontera entre las regiones, puede ser necesario ajustar una línea a esos puntos. Esto conlleva un incremento en el tiempo de ejecución y una ineficiencia computacional evidente, especialmente si hay muchos puntos de bordes. Un método para encontrar tales líneas de frontera es la transformada de Hough.

La Transformada de Hough fue diseñada originalmente para encontrar líneas en las imágenes [44], pero esta puede ser fácilmente variada para encontrar otras formas. Ésta es un tipo de transformación, en la cual el objeto de la imagen es expresado con una función matemática de algunos parámetros y el mismo, puede ser representado en otro dominio por medio de ellos. Por ejemplo, si se considera un contorno recto, en el domino de representación original se puede usar $y = ax+b$ para describir este borde, con una pendiente a y una intersección b, como muestra la Figura 12(a). Con el fin de desarrollar los dos parámetros a y b, se puede convertir el problema al dominio $a – b$ a partir del espacio original $x – y$, y tratar a x y y como parámetros. De esta manera, se encuentra que para cada punto (x, y) sobre el contorno, hay un número infinito de puntos correspondientes

Figura 12. Representación de la transformada de Hough



(a) Espacio *x - y* (b) Espacio *a - b*

Fuente: el autor

(a, b) y forman una línea en el espacio *a – b, b = xa + y.* Entonces, las líneas que son producidas por puntos sobre los bordes (rectos), idealmente deberán interceptarse en un solo punto, el cual indica la verdadera pendiente e intercepción del contorno, como ilustra la Figura 12(b). Para localizar este punto de intersección en el espacio paramétrico, se crea una *matriz acumuladora*, en la cual el número de veces que un cierto pixel es intersecado en el espacio paramétrico por una línea (o cualquier curva) es tratado como la intensidad del punto. Así, el valor adecuado de los parámetros se puede calcular a partir de las coordenadas del pixel más brillante en el espacio paramétrico.

La transformada de Hough también puede usarse para detectar parábolas [45]. En este caso, los parámetros que describen las curvas se pueden calcular en el espacio paramétrico 3D de $\mathrm{h}-\mathrm{k}-\mathrm{p}$ mediante tres parámetros, según la expresión:

$$(y-k)^2 = -4p(x-h)^2 \qquad (1.3)$$

Cada punto de la parábola en el espacio $x-y$ corresponde a la parábola en el espacio $h-k-p$. Todos los puntos sobre la parábola en el espacio $(x-y)$ se intersecan en un punto en el espacio $h-k-p$.

Para implementar la transformada de Hough parabólica es necesario crear celdas de acumulación, que consisten en celdas matriciales que se inicializan en cero y donde se almacenan el número de aciertos de intersección de las parábolas.

1.3.6 Descripción

El procesamiento digital de imágenes termográficas, puede ser muy útil como técnica de diagnóstico temprano de enfermedades, como cáncer de mama, tumores malignos, enfermedades en el ojo, etc. Existen numerosos métodos para describir las características de una imagen. Un amplio grupo de métodos están basados en el cálculo de parámetros estadísticos, como la media, desviación estándar, asimetría, curtosis, entropía, etc. y pueden utilizarse para comparar las imágenes térmicas. Los parámetros estadísticos de primer orden usan el histograma de la imagen para calcular sus valores, mientras los descriptores de segundo orden son obtenidos a partir de la matriz de co-ocurrencia de la imagen.

1.3.6.1 Descriptores estadísticos de primer orden

El histograma es una función que muestra para cada valor de pixel *i*, el número de pixeles *H(i)* (proporcional a la frecuencia) que en la imagen tienen ese valor. En la forma gráfica del histograma, *i* se representa por el eje horizontal, y *H(i)* por el vertical. La Figura 9c es un ejemplo de histograma. Puesto que la imagen es una matriz rectangular discretizada, los pixeles toman el valor *i* en un rango *[0, L-1]* (*L* no es necesariamente el número de valores que pueden tomar los pixeles, más bien, es el número de contenedores, bajo los cuales se calculan las frecuencias del histograma). El histograma también se puede interpretar como la función de densidad de probabilidad. Por ejemplo *p(i=3) = H(3)/16*, es la probabilidad de que un pixel, que Pertenece a una imagen de 4x4 elementos, tenga valor 3. Es decir, que *p(i)* es la densidad de probabilidad, calculada como la frecuencia del nivel respectivo, dividida entre el número de contenedores (*L*).

Tabla 2. Descriptores estadísticos de primer orden

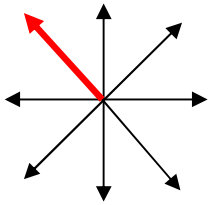
Nombre	**Formula**
Media	$\mu = \sum_{i=0}^{L-1} ip(i)$
Desviación estándar	$\sigma = \sum_{i=0}^{L-1} (i-\mu)^2 \, p(i)$
Asimetría	$\frac{\sum_{i=0}^{L-1} (i-\mu)^3 \, p(i)}{\sigma^3}$
Curtosis	$\frac{\sum_{i=0}^{L-1} (i-\mu)^4 \, p(i)}{\sigma^4}$
Energía	$\sum_{i=0}^{L-1} p^2(i)$
Entropía	$\sum_{i=0}^{L-1} \log_2\big(p(i)\big) p(i)$
Moda	Es el dato que aparece con mayor frecuencia
Mediana	Datos menores o iguales que la mediana representarán el 50% de los datos, y los que sean mayores representarán el otro 50%.
Máximo	Es el valor que tiene la mayor magnitud
Rango	Es la diferencia entre el máximo y el mínimo

En la Tabla 2 se muestran los parámetros estadísticos de primer orden: media, desviación estándar, asimetría, curtosis, energía y entropía, obtenidos a partir del histograma. También se exponen los descriptores: máximo, mediana, moda y rango, no necesariamente definidos a partir del histograma.

1.3.6.2 Descriptores estadísticos de segundo orden

Los descriptores de segundo orden pueden proveer más información de las imágenes termográficas que se desean analizar. Estos parámetros son definidos a partir de la matriz de co-ocurrencia [46]. Este arreglo matricial, representa la

Figura 13. Ejemplo de creación de la matriz de co-ocurrencia, con distancia de 1 pixel



(a) Dirección del ejemplo

1	0	2	0
1	1	0	2
2	1	0	2
1	1	2	0

(b) Matriz original de datos

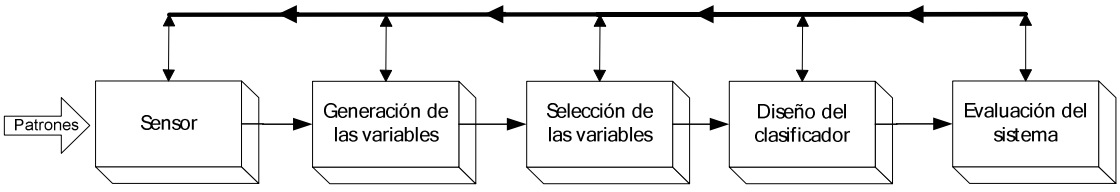
2	1	0
0	2	1
1	1	1

(a) Matriz de co-ocurrencia

Fuente: el autor

probabilidad conjunta de que dos pixeles tengan un valor de intensidad *"i"* y *"j"*, respectivamente, a una distancia *d*, en una dirección determinada. La matriz de co-ocurrencia considera no sólo la información sobre los niveles de intensidad, sino también la posición de los pixeles con valores de intensidad similares. En la Figura 13 se muestra una matriz de 4 x 4 con tres niveles de intensidad. Al lado, se muestra la matriz de co-ocurrencia, con distancia *d = 1* y dirección diagonal que apunta hacia el noroeste. Su dimensión es 3 x 3, pues los niveles de intensidad son: 0, 1 y 2. El elemento en la posición (1,1) de la matriz de co-ocurrencia, indica que el nivel "0" esta junto al nivel "0" 2 veces, en la dirección diagonal noroeste. El elemento en la posición (2,3), relaciona el nivel "1" junto al nivel "2" en la misma dirección, presentándose una vez. Ambos casos se resaltan con los colores azul y verde, respectivamente.

Figura 14. Etapas fundamentales en el diseño de un sistema de clasificación



Fuente: adaptado de [47]

Tabla 3. Descriptores de segundo orden, obtenidos con la matriz de co-ocurrencia

Nombre	Formula
Segundo momento angular, uniformidad o energía	$f_1 = \sum_i \sum_j p(i,j)^2$
Contraste	$f_2 = \sum_{n=0}^{N-1} n^2 \left\{ \sum_{i=1}^{N} \sum_{\substack{j=1 \\ \lvert i-j \rvert = n}}^{N} p(i,j) \right\}$
Correlación	$f_3 = \dfrac{\sum_i \sum_j (ij)\, p(i,j) - \mu_x \mu_y}{\sigma_x \sigma_y}$
Varianza	$f_4 = \sum_i \sum_j (i-\mu)^2 \, p(i,j)$
Inversa del momento de la diferencia	$f_5 = \sum_i \sum_j \dfrac{1}{1+(1-j)^2} \, p(i,j)$
Promedio de la suma	$f_6 = \sum_{k=2}^{2N} k p_{x+y}(k)$
Varianza de la suma	$f_7 = \sum_{k=2}^{2N} \left\{ (k - f_6)^2 \, p_{x+y}(k) \right\}$
Entropía de la suma	$f_8 = \sum_{k=2}^{2N} p_{x+y}(k) \log\left\{ p_{x+y}(k) \right\}$
Entropía	$f_9 = -\sum_i \sum_j p(i,j) \log\left\{ p(i,j) \right\}$
Varianza de la diferencia	$f_{10} = \sum_{k=0}^{N-1} \left[k - \sum_{l=0}^{N-1} l p_{x-y}(l) \right]^2 p_{x-y}(k)$
Entropía de la diferencia	$f_{11} = -\sum_{k=0}^{N-1} p_{x-y}(k) \log\left\{ p_{x-y}(k) \right\}$
Medida de correlación 1	$f_{12} = \dfrac{f_9 - HXY_1}{\max(HX, HY)}$
Medida de correlación 2	$f_{13} = \sqrt{1 - \exp\left[-2\left(HXY_2 - f_9 \right) \right]}$
Coeficiente de máxima correlación	$f_{14} = \sqrt{\text{segundo mayor valor propio de Q}}$
Homogeneidad	$f_{15} = \sum_i \sum_j \dfrac{1}{1+\lvert 1-j \rvert} \, p(i,j)$
Máxima probabilidad	$f_{16} = \max_{i,j} p(i,j)$

$p(i,j)$ es el valor del elemento *(i, j)* en la matriz de co-ocurrencia; $\sum_i = \sum_{i=1}^{N}$; $\sum_j = \sum_{j=1}^{N}$

N es la dimensión de la matriz de co-ocurrencia (N x N). μ es la media de μ_x y μ_y

$p_x(i) = \sum_{j=1}^{N} p(i,j)$; $p_y(j) = \sum_{i=1}^{N} p(i,j)$; $\mu_x = \sum_{i=1}^{N} i p_x(i)$; $\mu_y = \sum_{j=1}^{N} j p_y(j)$

$\sigma_x = \sqrt{\sum_{i=1}^{N} p_x(i)(i-\mu_x)^2}$; $\sigma_y = \sqrt{\sum_{j=1}^{N} p_y(j)(j-\mu_y)^2}$; $Q(i,j) = \sum_{k=1}^{N} \dfrac{p(i,k)\, p(j,k)}{p_x(i)\, p_y(j)}$

$p_{x+y}(k) = \sum_{i=1}^{N} \sum_{\substack{j=1 \\ i+j=k}}^{N} p(i,j)$; $k = 2,3,\ldots 2N$; $p_{x-y}(k) = \sum_{i=1}^{N} \sum_{\substack{j=1 \\ \lvert i-j \rvert = k}}^{N} p(i,j)$; $k = 0,1,\ldots N-1$

$HXY_1 = -\sum_{i=1}^{N} \sum_{j=1}^{N} p(i,j) \log\left\{ p_x(i)\, p_y(j) \right\}$; $HXY_2 = -\sum_{i=1}^{N} \sum_{j=1}^{N} p_x(i)\, p_y(j) \log\left\{ p_x(i)\, p_y(j) \right\}$

HX y *HY* son las entropías de p_x y p_y, respectivamente.

En la Tabla 3 se definen varios descriptores de segundo orden, a partir de la matriz de co – ocurrencia [40][46].

1.3.7 Reconocimiento de Patrones

El reconocimiento de patrones busca clasificar objetos en un número de categorías o clases. Dependiendo de la aplicación, estos objetos pueden variar desde señales unidimensionales, imágenes o cualquier tipo de medida que necesite ser clasificada. Las aplicaciones que usan el reconocimiento de patrones son variadas: reconocimiento de voz, caracteres, rostros, huellas, aplicaciones en robótica, diagnóstico mediante ayuda computacional (CAD[11]), etc. El CAD es de particular interés, puesto que ayuda a los médicos a tomar decisiones respecto al diagnóstico de enfermedades. La ayuda diagnóstica ha sido aplicada en una variedad de datos clínicos, como: rayos X, tomografías, imágenes ultrasónicas, electrocardiogramas, electroencefalogramas, termografía infrarroja, entre otras.

La Figura 14 muestra las diferentes fases en el diseño de un sistema clasificador. Estas etapas, se pueden describir mediante las siguientes preguntas [47]:

- ¿Cómo se obtiene la información?
- ¿Cómo se generan las variables?
- ¿Cuál es el mejor número de variables a usar?
- ¿Cómo se debe diseñar el clasificador?
- ¿Qué tan efectivo es el clasificador diseñado?

Existe un amplio número de algoritmos que pueden ser utilizados en el reconocimiento de patrones: clasificadores de Bayes, redes neuronales, máquinas de soporte vectorial, algoritmos genéticos, lógica fuzzy, etc.

[11] Computer-aided diagnosis

1.4 INTELIGENCIA ARTIFICIAL PARA EL DIAGNÓSTICO MÉDICO

El incremento de los conocimientos y el desarrollo de nuevas técnicas diagnósticas, han hecho cada vez más compleja la toma de decisiones por parte de los clínicos, que han continuado realizando el análisis de la información de forma manual y subjetiva. Por esto, el uso de sistemas informáticos se ha convertido en una de las áreas de gran interés en la medicina.

La informática médica se ha venido aplicando desde finales del siglo XIX, pero es a finales de los años 60 que se inicia el desarrollo de métodos de ayuda diagnóstica basados en Inteligencia Artificial (IA) [49]. Los primeros sistemas médicos fueron desarrollados por universidades de Estados Unidos, los cuales mostraban la posibilidad de realizar diagnósticos de diferentes patologías [48]: en Rutgers University se diseñó CASNET[12] para el diagnóstico y tratamiento de enfermedades. A mediados de 1970, en Standford se desarrolló MYCIN, para diagnosticar infecciones microbianas y dar una recomendación de tratamiento con medicamentos. El Massachusetts Institute of Technology diseñó en 1970, PIP[13], un simulador para la clasificación de la historia médica de un paciente dentro de un grupo de enfermedades renales conocidas. En 1980 se desarrolla en Pittsburg, INTERNIST, para el diagnóstico de medicina interna a través de la investigación de métodos heurísticos, utilizando el diagnóstico diferencial en la toma de decisiones clínicas.

Posteriormente en los años 90 aparecieron nuevos modelos, entre ellos las Redes Neuronales Artificiales (RNA), las cuales han tenido gran éxito en el desarrollo de ayudas médicas. Se destacan aplicaciones sobre tratamientos de la arteria coronaria, infartos miocárdicos, diagnóstico de cáncer, diagnóstico de pulmonía y desórdenes cerebrales, entre otros [49].

---

[12] Causal Associational Networks

[13] Present Illmes Program

Las aplicaciones de apoyo para el diagnóstico de cáncer de mama se han venido estudiando e implementando, sirviendo de complemento a métodos tradicionales como la mamografía y la termografía, logrando resultados interesantes. En 1992, Wu et al. [50] aplicaron RNA a los resultados de mamografías digitales para detectar microcalcificaciones agrupadas y obtuvieron una reducción del 50% en falsos positivos, manteniendo el 95% de verdaderos positivos. Lo J. Y. et al. [51] evaluaron la contribución de variables de la historia médica a la predicción de cáncer de mama con RNA y resultados de mamografías, encontrando que para una especificidad dada de 98%, adicionar la edad a la RNA entrenada con los hallazgos mamográficos, aumentaba la sensibilidad de 39% a 42%. Para esta misma especificidad la impresión del radiólogo había obtenido una sensibilidad sólo del 12%. Ng. Et al. [52] evaluaron una técnica que integra termografía, RNA y métodos bio-estadísticos, la cual obtuvo una tasa de precisión de 80.95%, con 100% de sensibilidad y 70.6% de especificidad en la identificación de cáncer de mama. Los resultados son exitosos al compararlos con el examen clínico que obtuvo una tasa de precisión entre el 60% y el 70%. En el análisis de termografía de mama con RNA, Koay J. et al. [13], demuestran que una RNA BP[14] fue capaz de generar una salida bastante precisa, al ser entrenada con parámetros estadísticos del termograma.

1.4.1 Redes Neuronales Artificiales (RNA)

Las RNA son modelos que buscan emular las habilidades del cerebro, de forma que mediante entrenamiento, llevan a cabo un proceso de aprendizaje para ejecutar tareas específicas. Están compuestas por unidades simples de procesamiento o *neuronas*, las cuales se encuentran interconectadas entre sí, formando una estructura para la transmisión de la información y operando de forma paralela. A cada una de estas conexiones se asocia un *peso sináptico* en el cual se representa el conocimiento [53]. Entre las propiedades del cerebro biológico que se observan en las RNA se destacan: la habilidad de adaptación,

[14] Backpropagation

generalización y la tolerancia a fallos, gracias a su procesamiento paralelo.

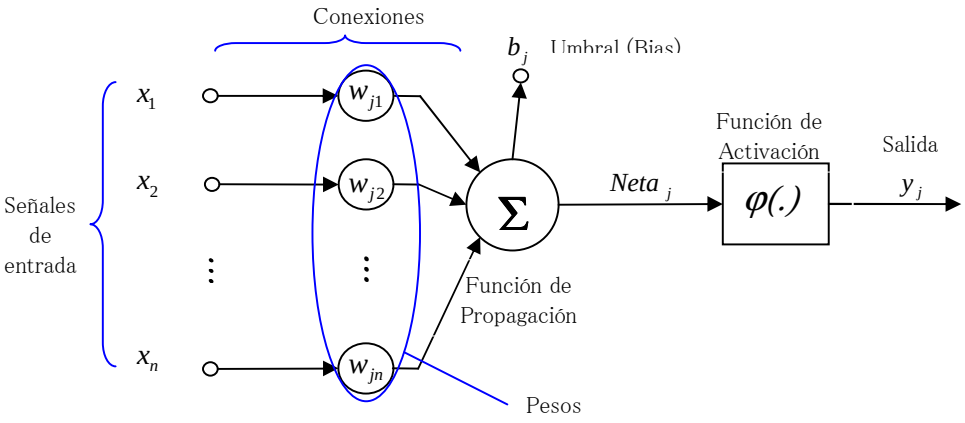
1.4.1.1 Modelo general de una neurona artificial

Entre los elementos básicos que conforman la estructura de neurona más utilizada (Figura 15), se encuentran los siguientes [53][54]:

- Entradas: x_i.
- Conexiones o sinapsis: asociadas a un peso sináptico, el cual indica la intensidad de interacción entre la neurona presináptica *i* y la neurona postsináptica *j*. Dependiendo del signo, el peso sináptico es excitatorio cuando es positivo, inhibitorio cuando es negativo, o tomar valor 0 cuando no existe conexión.
- Función o regla de propagación: calcula el valor del potencial post-sináptico de la neurona conocido como *Neta*. Por lo general se utiliza como función de propagación, la suma ponderada de las señales de entrada por sus respectivos pesos sinápticos, incluyendo su *umbral* o *Bias*, denotado por b_i, si existiera. Se puede describir como:

$$Neta_j = \sum_i^n w_{ji} x_i + b_i \qquad (1.4)$$

Figura 15. Modelo de una neurona



Fuente: Adaptación de [53]

En donde x_i son las señales de entrada provenientes de otras neuronas, w_{ij} son los pesos sinápticos de la neurona *j*, b_i es el umbral y *n* el número de señales de entrada.

- Función de Activación: representa el estado actual de la neurona *j.*

$$y_j = \varphi(Neta_j) \qquad (1.5)$$

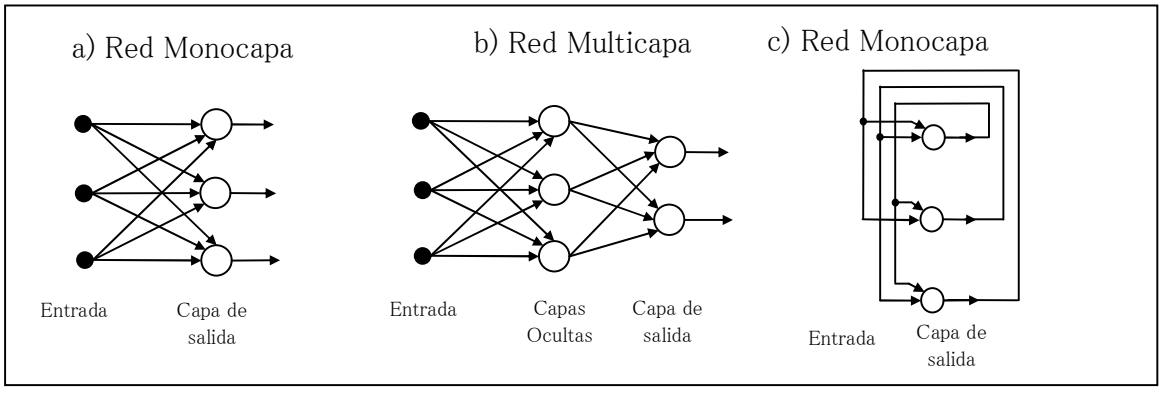
Las funciones de activación comúnmente utilizadas en aplicaciones prácticas son: función Identidad, Escalón, Lineal a tramos, Sigmoidea, Gausiana, entre otras.

- Función de salida: ocasionalmente se tiene en cuenta cuando ésta es diferente a la función identidad, la cual es la utilizada con mayor frecuencia.

1.4.1.2 Arquitecturas de las redes neuronales

La arquitectura se refiere, a la disposición de las conexiones entre neuronas dentro de la red. Por lo general, las neuronas se agrupan unidades denominadas *capas*. Dentro de una capa, las neuronas son del mismo tipo y pueden formar *grupos neuronales*. La agrupación de una o más capas conforma la *red neuronal* [54]. Según el número de capas, las redes neuronales se pueden clasificar en: redes monocapa y Multicapa (Figura 16).

Figura 16. Ejemplos de arquitecturas neurales



Fuente: adaptado de [53]

De acuerdo a la dirección del flujo de datos dentro de la red se pueden clasificar en: redes unidireccionales (*feedforward*) y redes recurrentes (*feedback*).

1.4.1.3 Tipos de aprendizaje

El proceso de aprendizaje también conocido como *algoritmo de aprendizaje*, se realiza a través de la creación o modificación de los pesos sinápticos, con el fin de obtener la salida deseada. Entre los más conocidos se encuentran [54]:

- Aprendizaje supervisado: consiste en ingresar un grupo de patrones a la red junto a la salida deseada. La red ajusta los pesos de las neuronas, corrigiendo el error, de forma que al presentarle posteriormente esos patrones los relacione con la salida memorizada. Se encuentran en este grupo el Perceptrón, Adaline, *Backpropagation*, entre otros.

- Aprendizaje no supervisado o autoorganizado: se presentan los patrones a la red y ésta los agrupa de acuerdo a rasgos similares. Por ejemplo los Mapas de Kohonen y Neocognitrón.

- Aprendizaje híbrido: combina el aprendizaje supervisado y el autoorganizado, los cuales se presentan en diferentes capas de neuronas. Entre ellos se destaca las funciones de base radial (Radial Basis Function RBF).

- Aprendizaje reforzado: Se encuentra entre el aprendizaje supervisado y el autoorganizado. Determina un error que indica el rendimiento global de la red pero no se le proporciona la salida deseada. Ejemplo de este tipo de aprendizaje es el Premio-castigo asociativo.

1.4.1.4 Backpropagation (BP)

El algoritmo BP utiliza el método de gradiente descendiente para minimizar el error entre la salida actual y la deseada. Se basa en el ciclo “propagación-adaptación” para el aprendizaje de un conjunto de patrones de entrada-salida previamente establecidos. Este consta de dos fases [55]:

- La primera fase es el aprendizaje “hacia adelante”, en donde se introducen los patrones de entrada, los cuales se propagan a través de las capas ocultas hasta producir una salida.
- En la segunda fase o aprendizaje “hacia atrás”, se realiza una comparación entre la salida actual y la deseada para calcular el error de cada neurona de la capa de salida. Este error se propaga desde la última capa a la inmediatamente anterior y así a través de toda la red, con el fin reajustar los pesos y minimizar el error en el futuro.

1.4.1.5 Radial Basis Function (RBF)

En la arquitectura RBF, las entradas envían información del exterior hacia las capas ocultas, en donde es procesada y transmitida hacia la capa de salida conformada por neuronas de tipo lineal. En esta capa se calcula la suma ponderada y se obtiene la salida final de la red [55]. El proceso en las capas ocultas es realizado por funciones de base radial, las cuales calculan la distancia entre las entradas y el vector de pesos asociados a cada neurona, denominado centroide, al cual se le aplica una función radial gaussiana. Este tipo de aprendizaje se realiza en dos fases [55]:

- En la primera fase se introducen todos los patrones de entrada para entrenar los grupos neuronales de las capas ocultas. Estos actualizan su centroide a través de iteraciones sucesivas hasta la convergencia.

- Finalmente, en la segunda fase se ajustan los pesos de las conexiones de la capa de salida

1.4.1.6 Algoritmo de Correlación en Cascada (CC)

La arquitectura de Correlación en Cascada representa un tipo de algoritmo compuesto, en el cual, algunas técnicas de aprendizaje se encuentran incorporadas (Backpropagation, Resilient Backpropagation, etc.). CC se caracteriza por ser un algoritmo de aprendizaje constructivo. Este comienza con una red mínima, compuesta de una capa de entrada y una capa de salida, a partir de la cual se añaden neuronas a la capa oculta, paso a paso, minimizando el error de toda la red [56].

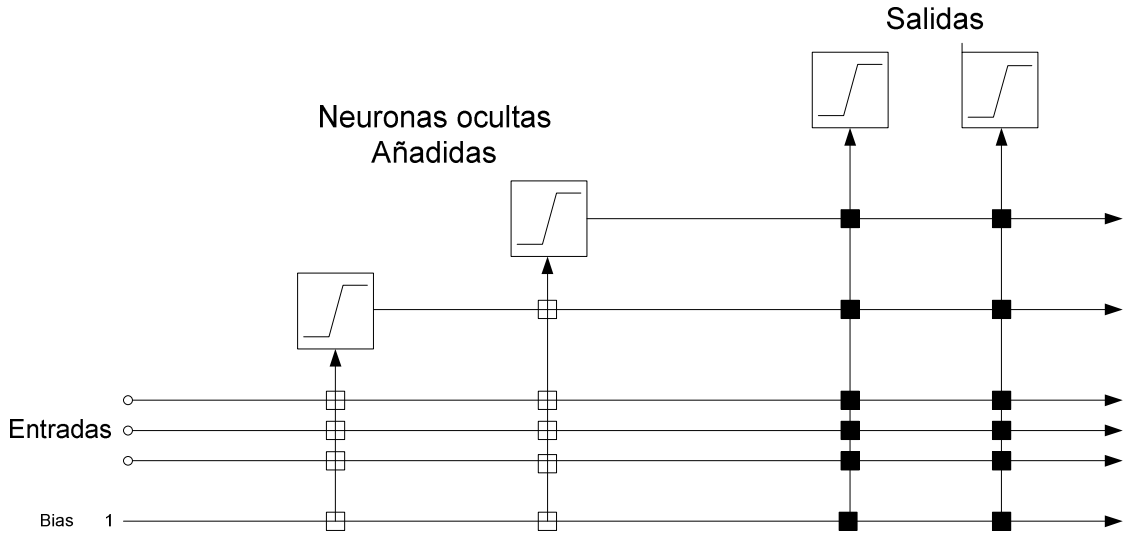
Correlación en Cascada es una arquitectura con aprendizaje supervisado, en la cual se construye una topología multicapa mínima. Las dos ventajas de esta arquitectura son:

- No es necesario que el usuario se preocupe por la cantidad de neuronas y su disposición, y
- CC es mucho más rápido que los algoritmos de entrenamiento usuales.

La Correlación en Cascada combina dos ideas: la primera es la arquitectura en cascada, en la cual las neuronas ocultas son añadidas una a la vez y no cambian luego de ser adheridas. La segunda es la técnica de aprendizaje; esta crea e instala las nuevas neuronas ocultas. Para cada nueva unidad oculta, el algoritmo maximiza la magnitud de la correlación entre la salida de la nueva neurona y el error residual de la red.

El algoritmo se realiza de la siguiente forma:

Figura 17. Una red neuronal entrenada con CC después de que 2 neuronas ocultas han sido agregadas. Las líneas verticales llevan todas las activaciones entrantes. Las conexiones con cuadros blancos están congeladas. Las conexiones con cuadros negros son entrenadas repetidamente.



Fuente: adaptado de [56]

1. CC inicia con una red mínima que consiste de una capa de entrada y una de salida.
2. Se entrenan todas las conexiones que terminan en una neurona de salida con un algoritmo típico, hasta que el error de la red no disminuye más.
3. Se generan las *neuronas candidatas*. Cada unidad candidato es conectada con todas las neuronas de entrada y con todas las neuronas ocultas existentes. Entre las unidades ocultas y de salidas no hay pesos.
4. Se intenta maximizar la correlación entre la activación de las neuronas candidatas y el error residual de la red, entrenando todas las conexiones que llegan a la unidad candidata. El entrenamiento se detiene cuando la correlación no mejora.
5. Se elige la neurona candidata con la máxima correlación, se congelan sus pesos entrantes y se agregan a la red. Para cambiar la candidata a una

nueva neurona oculta, se generan conexiones entre la unidad seleccionada y todas las neuronas de salida. Puesto que, los pesos que llegan a la nueva neurona se mantienen, un nuevo detector de características es obtenido. Se regresa al paso 2.

Este algoritmo se repite hasta que todo el error de la red yace por debajo de un valor determinado. La Figura 17 muestra una red después de que dos neuronas ocultas han sido añadidas.

El entrenamiento de las neuronas de salida intenta minimizar la suma cuadrática de los errores:

$$E = \sum_{p} \frac{1}{2} \sum_{o} (y_{po} - t_{po})^2$$

Donde t_{po} es el valor deseado y y_{po} es la valor observado de la neurona de salida *o* para un patrón *p*. El error *E* es minimizado mediante el gradiente descendiente, usando:

$$e_{po} = (y_{po} - t_{po}) f'_{p}(net_{o})$$

$$\frac{\partial E}{\partial \omega_{io}} = \sum_{p} e_{po} I_{ip}$$

Donde f'_{p} es la derivada de una función de activación de una neurona de salida *o*, además I_{ip} es el valor de una unidad de entrada o una neurona de salida *i* para un patrón *p*. ω_{io} es el peso de la conexión entre neurona de entrada o neurona de salida *i* y una unidad de salida *o*.

Después de la fase de entrenamiento, las neuronas candidatas son adaptadas, tal que la correlación[15] *C* entre el valor y_{po} de una candidata y el error residual e_{po} de una neurona de salida, sea máxima. La correlación esta dada por [56]:

[15] En realidad es una covarianza

$$C = \sum_{o} \sum_{p} \left| (y_{po} - \bar{y}_o)(e_{po} - \bar{e}_o) \right|$$

Donde $\bar{y}_o$ es la activación promedio de una neurona candidata y $\bar{e}_o$ es el error promedio de una neurona de salida, sobre todos los patrones *p*.

1.4.2 Fuzzy C- means clustering

La técnica, conocida con el nombre en inglés de *Fuzzy C-means* (FCM), es un algoritmo de agrupamiento, en el cual cada punto de los datos pertenece a un grupo (cluster) con cierto grado, especificado por un valor de membresía [57].

FCM divide una colección de n vectores $\boldsymbol{x_i}$, con $i = 1, 2, \ldots, n$, en c grupos *fuzzy* y encuentra un centro en cada clúster, de forma que, una medida de disimilitud (función objetivo) sea minimizada. La principal diferencia entre FCM y el conocido algoritmo *Hard C – means* (HCM), es que FCM emplea particiones "fuzzy", tal que, un punto puede pertenecer a varios grupos con un grado de pertenencia especificado por una membresía entre 0 y 1. Así, la división de los grupos es definida a través de una matriz de membresía **U** de $c \times n$, donde el elemento u_{ij} tiene valores entre 0 y 1. Sin embargo, se impone una normalización para que la suma de los grados de pertenencia para un conjunto de datos siempre sea igual a la unidad:

$$\sum_{i=1}^{c} u_{ij} = 1, \quad \forall j = 1, 2..., n \qquad (1.6)$$

La función objetivo para FCM es:

$$J(U, c_1, ...c_c) = \sum_{i=1}^{c} J_i = \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}^m d_{ij}^2 \quad (1.7)$$

Donde u_{ij} está entre 0 y 1, c_i es el centro del cluster del grupo i fuzzy, $d_{ij} =$

$\|\boldsymbol{c_i} - \boldsymbol{x_i}\|$ es la distancia Euclidiana entre el i-ésimo centro del grupo y el j-ésimo punto y m es un exponente de ponderación entre $[1, \infty)$.

Las condiciones necesarias para minimizar la ecuación (1.7), obtenidas mediante cálculo diferencial, son:

$$c_i = \frac{\sum_{j=1}^{n} u_{ij}^{m} \boldsymbol{x}_j}{\sum_{j=1}^{n} u_{ij}^{m}} \qquad (1.8)$$

Y

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{d_{ij}}{d_{kj}} \right)^{2/(m-1)}} \qquad (1.9)$$

El algoritmo de fuzzy C-means es un procedimiento de iteración a través de las dos ecuaciones anteriores. En modo de bloque, FCM determina los centros de los clúster $\boldsymbol{c_i}$ y la matriz de membresía **U** usando los siguientes pasos [58][59]:

1. Inicia la matriz de membresía **U** con valores aleatorios entre 0 y 1, siempre que la ecuación (1.6) sea satisfecha.
2. Calcula los *c* centros de los cluster fuzzy $\boldsymbol{c_i} = 1, \ldots, c$, usando la ecuación (1.8).
3. Determina la función objetivo, de acuerdo a la ecuación (1.7). El algoritmo se detiene, si esta función se encuentra por debajo de un cierto valor de tolerancia, o si la mejora respecto a la iteración previa es inferior de un umbral.
4. Calcula una nueva matriz **U** usando la ecuación (1.9). Se repite el paso 2.

Los centros de los cluster pueden ser inicializados al comienzo y luego se empieza el proceso iterativo. No existe garantía para que FCM converja a una solución óptima. El rendimiento depende de los centros iníciales, así, otros algoritmos más rápidos se pueden usar para determinar los centros del inicio o se puede ejecutar varias veces FCM, comenzando cada vez con un conjunto diferente de centros.

1.5 CORRELACIÓN LINEAL

Las pruebas estadísticas mediante correlación lineal son empleadas con gran frecuencia en el campo de la bioestadística, la ingeniería, la economía, etc. A continuación se describe conceptualmente los coeficientes de Pearson y Spearman, utilizados especialmente para las variables por intervalos o de razón y para las variables categóricas, respectivamente.

1.5.1 Coeficiente de correlación de Pearson

Es una prueba estadística para analizar la relación entre dos variables medidas en un nivel por intervalos o de razón (variables numéricas). La prueba en sí no considera a una como independiente y a otra como dependiente, ya que no se trata de una prueba que evalúa la causalidad. El coeficiente de correlación de Pearson se define como:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2}\sqrt{\sum (y_i - \bar{y})^2}} \qquad (1.10)$$

El *signo indica la dirección de la correlación* (positiva o negativa) y el *valor numérico, la magnitud de la correlación.*

El coeficiente *r* puede asumir valores entre *-1* y *1*. El signo de la correlación, positivo o negativo, depende del numerador, puesto que el denominador esta compuesto de raíces cuadráticas positivas. Considerando el numerador, cuando las desviaciones de *x* y *y* varían de la misma forma, es decir, cuando *x* se encuentra distante de su media y el valor de *y* también lo hace en el mismo sentido, el producto del numerador es positivo, dado que, multiplicar dos números positivos o dos números negativos produce un valor positivo. Si por el contrario, las desviaciones varían de forma opuesta, por ejemplo, *x* se encuentra por encima de la media y *y* se encuentra por debajo de su media, entonces el producto del

numerador es negativo. Si las variaciones en *x* no cambian con un patrón lineal con las variaciones de *y*, entonces existe una asociación de desviaciones positivas y negativas que resultan en un coeficiente *r* cercano a cero.

El máximo valor de *r* es *+1,* el cual ocurre cuando todos los puntos yacen sobre una línea recta con pendiente positiva. El valor mínimo de *r* es *-1*, cuando todos los puntos yacen sobre una línea recta con pendiente negativa.

1.5.2 Coeficiente de correlación por rangos de Spearman

La prueba de correlación por rangos de Spearman es una alternativa no paramétrica al coeficiente de correlación de Pearson, cuando las medidas son valores ordinales en vez de niveles por intervalos o de razón. Este analiza la relación entre la variable *x* y la variable *y* en una muestra de *n* pares de datos (x_i, y_i). Las medidas individuales son ordenadas sobre cada variable y se determina el grado de correspondencia entre los pares organizados. A diferencia de la prueba de Pearson, que mide la fuerza de la correlación lineal, el coeficiente de Spearman sólo evalúa si los pares de datos tienen una relación *monótonamente creciente*, en la cual una variable se incrementa tanto como la otra aumente, o una relación *monótonamente decreciente*, en la cual una variable se decrementa tanto como la otra se incrementa.

El primer paso es asignar rangos, desde *1* a *n,* a las medidas de cada variable. Así, cada dato tiene una jerarquía sobre la variable *x* y sobre la variable *y*. Cuando existen rangos empatados, a cada uno se le asigna la media de los rangos que deberían tener si fuesen adyacentes pero no empatados. El coeficiente es positivo cuando la relación se incrementa monótonamente (+1 indica una perfecta correlación), negativo cuanto la relación se decrementa monótonamente (-1 indica una perfecta correlación) y cero cuando no existe relación. Hay dos maneras para calcular el coeficiente, dependiendo si hay o no

un empate en los rangos.

Mientras haya rangos empatados sobre la misma variable, el coeficiente de Spearman se calcula mediante la ecuación 1.10 del coeficiente de Pearson. Pero en este caso, x_i es el rango sobre una variable y y_i es el rango sobre la otra variable.

Cuando no hay empates en los rangos de cada variable, la ecuación se reduce a:

$$r = 1 - \frac{6\sum_{i}^{n} d_i^2}{n(n^2 - 1)} \qquad (1.11)$$

Donde n es el número de pares ordenados y d_i es la diferencia entre los dos rangos de cada par, obtenidos por la categorización individual sobre cada variable de forma separada, del más bajo al más alto y calculando las diferencias d entre los rangos de x y y.

El siguiente ejemplo muestra como calcular el coeficiente de Spearman, si no hay rangos empatados:

En un estudio se desea determinar la correlación lineal que existen entre la edad y el índice de masa corporal (IMC) en las mujeres. Los datos fueron recolectados a través de una encuesta llenada por 20 pacientes diferentes; la Tabla 4 presenta la información recopilada.

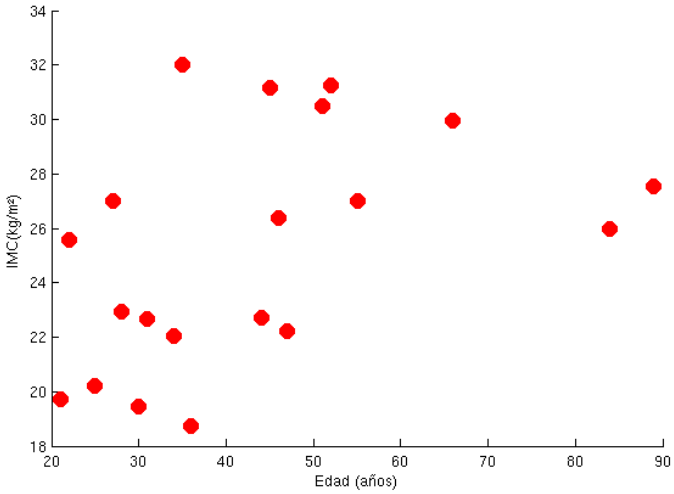
Para calcular el coeficiente de Spearman de la ecuación 1.11 se debe determinar la suma de todos los d_i, de la siguiente forma:

- Se ordenan los datos de la primera columa (edad) y se añade una nueva columna que corresponde a los rangos de esta variable (1, 2, 3, …,20).
- Se ordenan los datos de la segunda columna (IMC) y se añade una cuarta columna que asigna los rangos de esta variable de forma respectiva.

Tabla 4. Datos de la edad y el índice de masa corporal en un estudio para analizar la correlación entre estas variables.

Edad (años)	**IMC (kg/m^2)**
66	29,94
34	22,03
84	25,97
89	27,56
30	19,47
27	27,01
55	26,99
52	31,24
21	19,72
31	22,67
28	22,96
51	30,47
47	22,22
45	31,14
46	26,37
44	22,72
36	18,75
35	32,03
25	20,2
22	25,56

Figura 18. Gráfica de dispersión del IMC y la edad



Fuente: el autor

Tabla 5. Organización de los datos para el cálculo del coeficiente de Spearman

Edad (años)	**IMC (kg/m2)**	**Rango edad**	**Rango IMC**	***d***	***d*²**
21	19,72	1	3	-2	4
22	25,56	2	10	-8	64
25	20,2	3	4	-1	1
27	27,01	4	14	-10	100
28	22,96	5	9	-4	16
30	19,47	6	2	4	16
31	22,67	7	7	0	0
34	22,03	8	5	3	9
35	32,03	9	20	-11	121
36	18,75	10	1	9	81
44	22,72	11	8	3	9
45	31,14	12	18	-6	36
46	26,37	13	12	1	1
47	22,22	14	6	8	64
51	30,47	15	17	-2	4
52	31,24	16	19	-3	9
55	26,99	17	13	4	16
66	29,94	18	16	2	4
84	25,97	19	11	8	64
89	27,56	20	15	5	25

- Se crea una quinta columna con las diferencias *d* entre los rangos de la edad y el IMC.
- Finalmente, se adhiere una última columna con los valores del ítem anterior al cuadrado (d^2).

El resultado del proceso anterior se visualiza en la Tabla 5.

La suma de todos los valores de d^2 es 644. Utilizando la ecuación 1.11, el coeficiente se calcula como:

$$r = 1 - \frac{6(644)}{20(20^2 - 1)} = 0{,}5158$$

Este valor indica que existe una correlación positiva media, los cual se confirma a través de la Figura 18.

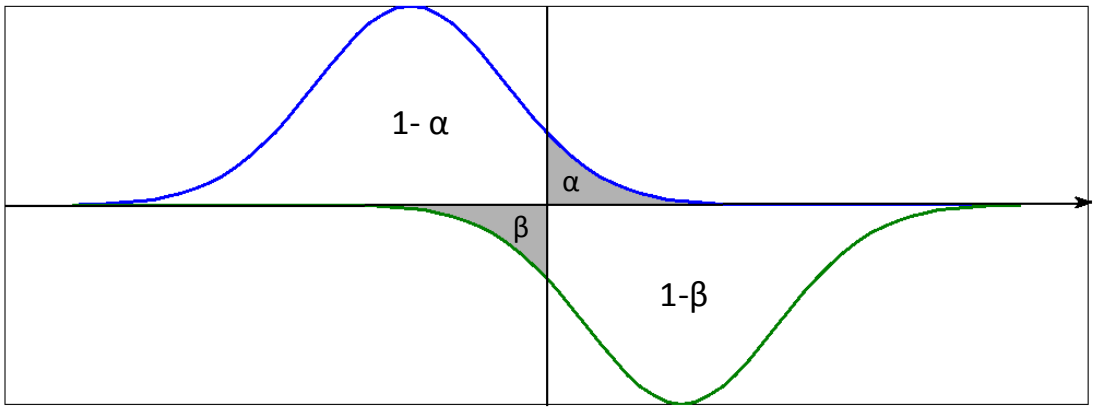
1.6 ANÁLISIS MEDIANTE LA CURVA ROC

La curva ROC (Receiver Operating Characteristic) es popular en la comunidad radiológica psicológica y médica, como medida que contiene diversa información en los estudios epidemiológicos. Además, la curva ROC se usa para juzgar la habilidad de selección de varios métodos estadísticos que combinan diversos resultados de pruebas, para propósitos predictivos. La representación e interpretación del área bajo la curva ROC puede obtenerse por predicciones matemáticas basadas en características presentadas por los pacientes, o por el método “rating”, tal como lo explica J. Hanley [60].

La Figura 19(a) ilustra un ejemplo de dos funciones de densidad de probabilidad (pdf[16]) que se traslapan, las cuales describen la distribución de una variable en dos clases, junto con un umbral (una de las pdf ha sido invertida para propósitos de ilustración). La primera clase representa los valores a la izquierda del umbral y la segunda clase representa los valores a la derecha. Esta decisión se asocia a un error de probabilidad, α, de alcanzar una decisión equivocada concerniente a la primera clase (la probabilidad de una correcta decisión es 1 – *α*). Esto es igual al área sombreada bajo la curva correspondiente. De igual forma, β es la probabilidad de una decisión equivocada (la decisión correcta es 1 – β), referente a la segunda clase. Moviendo el umbral sobre *todas* las posibles posiciones, se obtienen diferentes valores de *α* y β. Este proceso toma muy poco tiempo si las dos distribuciones de traslapan completamente, encontrándose que para cualquier posición del umbral α = 1 – β. Tal caso, corresponde a la línea recta diagonal en la Figura 19(b), donde los ejes son α y 1 – β.

[16] En inglés, probability density function

Figura 19. Ilustración para la interpretación de la curva ROC



(a) Ejemplo del traslape de dos pdf de la misma variable en dos clases

(b) Curva ROC resultante de varial el umbral en (a)

Fuente: adaptado de [47]

Tanto como las dos distribuciones se separen, la curva correspondiente se aleja desde la línea recta, como la Figura 19(b) demuestra. Si se analiza un poco más, se nota que entre menos traslape existe entre las clases, más grande es el área entre la curva y la línea. En el otro caso extremo, cuando las dos distribuciones de las clases están separadas totalmente, mover el umbral con un barrido sobre el

rango de valores de α entre 0 y 1, siempre produce valores de 1 – β igual a la unidad. Así, el área antes mencionada, varia entre cero, para un completo traslape y ½ (el área del triangulo superior), para una completa separación. Por esta razón, el área bajo la curva ROC *es una medida de la capacidad de discriminación entre clases de una variable específica*. En la práctica, la curva ROC puede ser fácilmente construida variando el umbral y calculando porcentajes de las clasificaciones equivocadas y correctas, sobre las variables disponibles.

La curva ROC, es un argumento gráfico que también se puede ver como la representación de *sensibilidad* versus *1 - especificidad*. Equivalentemente, la curva ROC puede ser representada, trazando la fracción de los aspectos verdaderos positivos (TPR = Tasa verdadera positiva) contra la fracción de aspectos falsos positivos (FPR = Tasa falsa positiva) [61].

Los parámetros de sensibilidad y especificidad se definen como:

$$Sensibilidad = \frac{VP}{VP + FN} \qquad Especificidad = \frac{VN}{VN + FP}$$

Donde, VP son los verdaderos positivos, FN los falsos negativos, VN los verdaderos negativos y FP los falsos positivos.

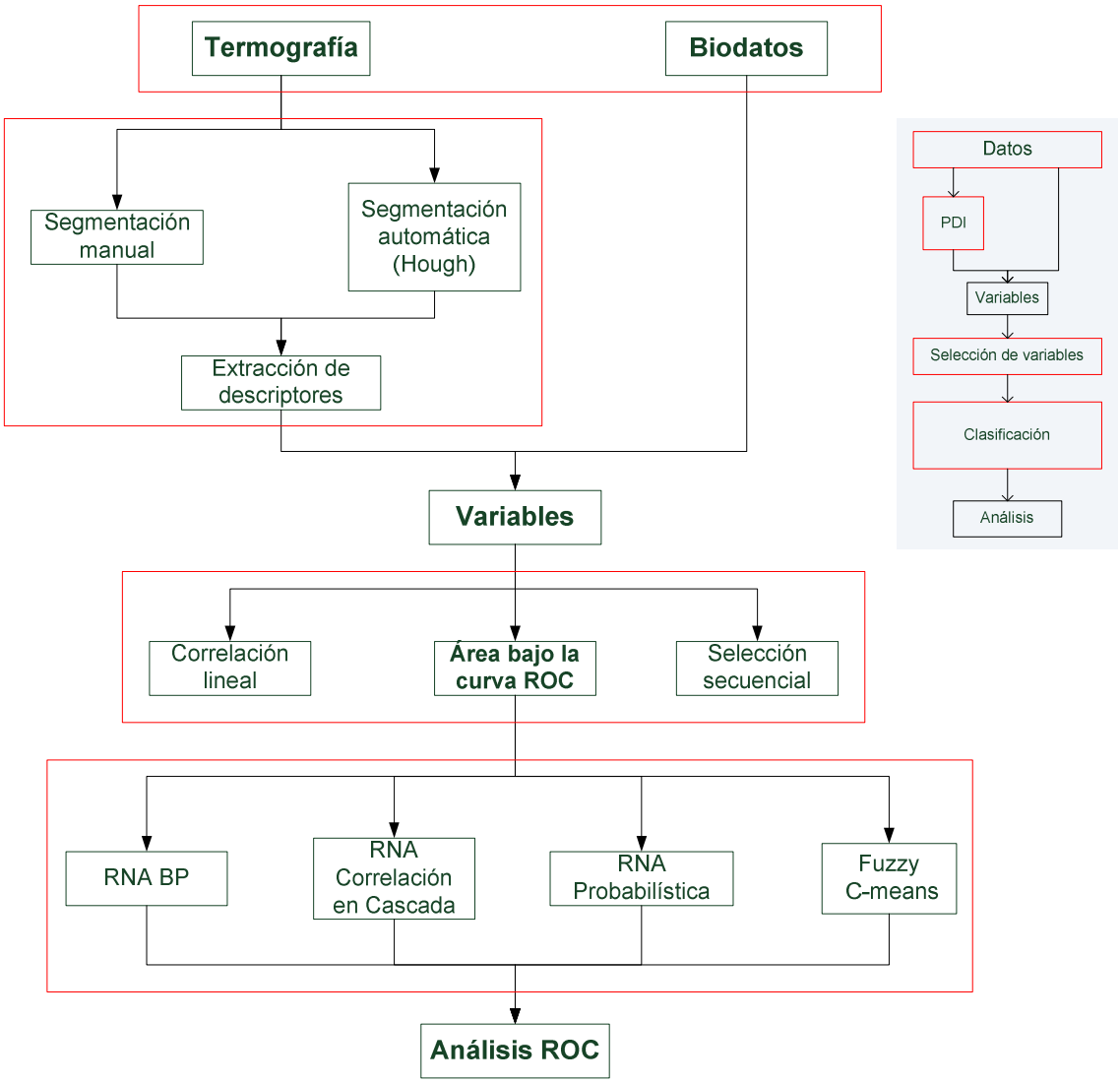
2 METODOLOGIA PARA LA DETECCIÓN DEL CARCINOMA DE GLÁNDULA MAMARIA FUSIONANDO VARIABLES TERMOGRÁFICAS Y CLINICAS

Este capítulo describe la metodología que se propone a lo largo del trabajo de investigación desarrollado. El objetivo principal de este estudio es clasificar una muestra poblacional de mujeres, en pacientes con cáncer de mama o ausencia del mismo, teniendo en cuenta la información termográfica de los senos y los datos clínicos de cada paciente.

La Figura 20 muestra la metodología que se plantea para la detección del carcinoma de mama. El diagrama de bloques a la derecha, representa el esquema general, el cual consta de: la adquisición de la información, el procesamiento digital de imágenes, la selección de las variables, la clasificación de las mismas y el análisis de los resultados. El diagrama a la izquierda, de mayor tamaño, describe de forma más detallada todo el proceso que se ha presentado.

El primer paso, fundamental en el desarrollo del proyecto, es la adquisición de la información, tanto térmica como clínica. Para esto, se desarrolla un protocolo de registro termográfico y preparación de la paciente, que incluyen todas las consideraciones bioéticas respectivas. Asimismo, se recolecta la información socio-demográfica, la historia familiar de cáncer, los antecedentes hormonales y los factores fisiológicos, a través de dos formularios, que la paciente personalmente llena. Esta fase, es la más difícil y delicada, pues se necesita una gran gestión médica para la participación y continua presencia de mujeres en el estudio. Una vez se han registrado de forma adecuada las termografías infrarrojas, se procede a elegir las regiones en la imagen que *contendrán* las glándulas mamarías. Es decir, el denominado proceso de segmentación, que se realiza de forma manual y automática.

Figura 20. Metodología para la detección del carcinoma de glándula mamaria fusionando variables clínicas y termográficas



Fuente: el autor

En la primera, se escoge la región del seno por medio de una elipse, alargada y rotada en dirección del mentón. La segunda clase de segmentación, se realiza principalmente a través de la unión del contorno del torso y la representación geométrica de la mama, en este caso, parabólica. Esta forma se detecta a través de la transformada parabólica de Hough.

Con las zonas plenamente designadas, de acuerdo a los dos métodos, se extraen los descriptores que posiblemente identifiquen las anomalías referentes a la presencia o ausencia de la enfermedad, los cuales, pueden ser descriptores de primer y segundo orden. Estos últimos, se calculan a partir de la matriz de co-ocurrencia, en 4 direcciones diferentes: horizontal, diagonal a 45º, vertical y diagonal a 135º. A la par, se preseleccionan los datos clínicos (y los térmicos), puesto que, algunas variables no tienen representación dado el número de muestra, o su interpretación conceptual se encuentra fuera del alcance de este trabajo.

Con los datos elegidos hasta el momento, se procede a realizar la selección de variables que posteriormente entraran al sistema de clasificación, según la relación existente con el resultado histopatológico. Se planean tres técnicas: la correlación lineal, que se realiza mediante los coeficientes de Pearson y Spearman. Este método es muy simple y eficaz, pero supone una independencia entre cada variable, lo cual no necesariamente ocurre. La selección secuencial de variables, que en este caso, adhiere parámetros al modelo, dependiendo de un criterio determinado. Y por último, el área bajo la curva ROC es usada como medida de separación entre clases de una variable. Esta técnica se complementa, al incluir un factor de ponderación que correlaciona la variable actual con las anteriormente escogidas. De esta forma, se emplea un método que mide tanto la relación entre variables, como la dependencia de estas con el diagnóstico de la biopsia. En realidad, el análisis del área bajo la curva ROC, es el único procedimiento que se aplica antes de la ejecución del algoritmo de inteligencia artificial. Sin embargo, las otras dos técnicas se implementan para un posterior análisis, con una muestra mucho más sustanciosa.

La selección definitiva se realiza para 8 conjuntos de datos, que resultan de las regiones segmentadas manual y automáticamente, a través de la combinación entre las variables clínicas, los descriptores de primer orden y los descriptores de

segundo orden; estos últimos, por cada una de las cuatro direcciones de la matriz de co-ocurrencia. A partir de aquí, se aplican cuatro diferentes algoritmos de clasificación para cada grupo de datos: una RNA backpropagation, una red neuronal con correlación en cascada, una RNA probabilística y la técnica de agrupamiento Fuzzy C-means. Para los tres primeros algoritmos, se entrena y prueba 15 veces con divisiones aleatorias de la muestra, para FCM se emplea toda la población, puesto que es un algoritmo no supervisado. Posteriormente, se determina la sensibilidad, especificidad y área bajo la curva ROC de cada prueba de clasificación y se comparan los resultados para cada tipo de algoritmo implementado.

3 ADQUISICIÓN DE DATOS

En este trabajo, los datos provienen de dos fuentes fundamentales: las termografías infrarrojas de los senos y, los datos socio-demográficos y clínicos de las pacientes involucradas en el estudio. En Colombia y en Latinoamérica, la termografía de mamas no se utiliza como técnica diagnóstica de enfermedades. Por esto, el esfuerzo involucrado en este proyecto es doble, ya que no existe antecedente en la región y se tiene que construir todo el proceso de adquisición, desde la asignación de la cita a la paciente hasta el registro de termografía infrarroja.

La cita para la obtención de los datos de cada paciente presenta cuatro fases:

- Información sobre el proyecto mediante el consentimiento informado
- Encuesta sobre los factores socio-demográficos, hereditarios y hormonales
- Acondicionamiento del paciente y registro termográfico
- Consulta clínica

3.1 CONSIDERACIONES BIOÉTICAS

Para tomar parte en el proyecto, los individuos deben ser participantes voluntarios e informados. Siempre debe respetarse el derecho de los participantes en la investigación a proteger su integridad y deben tomarse toda clase de precauciones para resguardar la intimidad de los individuos, la confidencialidad de la información del paciente y para reducir al mínimo las consecuencias de la investigación sobre su integridad física, mental y su personalidad.

Esta investigación se fundamenta en la declaración de Helsinki, que la Asociación Médica Mundial ha promulgado. Ésta es una propuesta de principios éticos, que

sirven para orientar los estudios médicos en seres humanos. Menciona la Declaración, que la investigación debe ejecutarse dentro del marco de la normatividad vigente en el país donde se realiza la investigación y que cada individuo participante debe recibir información adecuada acerca de los objetivos, métodos, fuentes de financiamiento, posibles conflictos de intereses, afiliaciones institucionales del investigador, beneficios calculados, riesgos previsibles e incomodidades derivadas. La persona debe ser informada del derecho de participar o no en la investigación y de retirar su consentimiento en cualquier momento, sin exponerse a represalias. Después de asegurarse que el individuo ha comprendido la información, el médico debe obtener entonces, preferiblemente por escrito, el consentimiento informado y voluntario de la persona.

En el ANEXO A se muestra el consentimiento informado que se elaboró en el proyecto macro. Este fue desarrollado junto al equipo médico y epidemiológico, adscritos a la investigación de COLCIENCIAS.

3.2 REGISTRO TERMOGRÁFICO

Para realizar el registro de las termografías de mama, se utilizó una cámara infrarroja Fluke® TI50, la cual tiene una resolución de 320 x 240, cubre un rango de temperaturas de -20 a +350 ºC, posee una sensibilidad térmica menor a 0.07 ºC, trabaja en la banda espectral de 8 a 14 µm y tiene una precisión térmica del 2%. Las especificaciones completas de la cámara se encuentran el ANEXO C.

La cámara infrarroja graba las termografías en formato propietario de Fluke® ".is2", pero el software distribuido por el fabricante permite exportar todas las temperaturas de la imagen en un archivo ".txt", que luego es convertido al formato .mat de matlab.

La recopilación de los termogramas debe estar gobernada por un protocolo de

registro para asegurar la calidad y el alcance de los análisis realizados. Los lineamientos establecidos en el protocolo condicionarán el escenario (laboratorio) del registro, los procedimientos y a las mismas pacientes que serán objeto del examen. El diseño del mismo se realizó con base en la experiencia adquirida en investigaciones anteriores en el campo de la termografía [62] y en las pruebas piloto que se realizaron durante las etapas iníciales de este proyecto.

3.2.1 Protocolo de manejo de la paciente

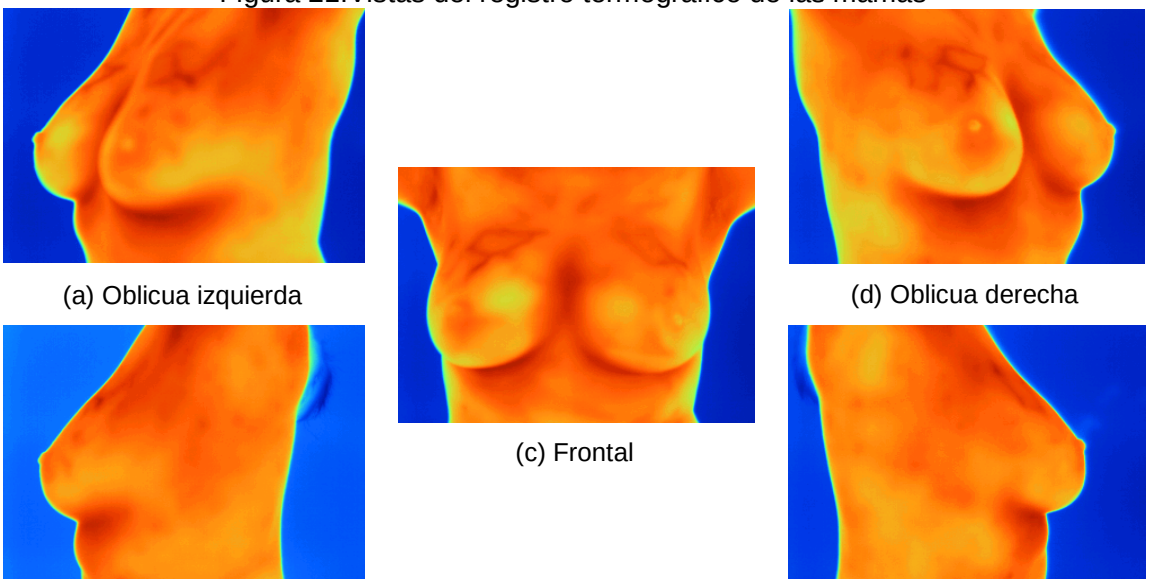
La temperatura superficial de la piel del ser humano es sensible a diversos factores como la transpiración, el metabolismo, la circulación, etc. Por lo tanto, se debe tener especial cuidado de la paciente antes y durante el registro de termografía. A continuación se mencionan algunos aspectos, respecto al manejo de la paciente:

1. No debió exponerse al sol (bronceado), 5 días antes del examen.
2. No debe usar lociones, cremas, polvos o maquillaje el día del examen.
3. No debe usar desodorante o antitranspirante el día del examen.
4. El área alrededor del seno debe estar depilada.
5. No debió haber realizado ninguna terapia física 24 horas antes del examen.
6. No debió haber realizado ejercicio mínimo 4 horas antes del examen.
7. Debe haberse bañado al menos una hora antes del examen.
8. No debe haber ingerido medicamentos para el dolor o vaso dilatadores el día del examen.
9. No debe usar ropa ajustada (camisa) el día del examen.

3.2.2 Protocolo de registro del termograma

Para la elaboración del registro de las termografías de mama, se deben atender los

Figura 21.Vistas del registro termográfico de las mamas



(a) Oblicua izquierda

(d) Oblicua derecha

(c) Frontal

(b) Lateral izquierda

(e) Lateral derecha

Fuente: el autor

siguientes ítems:

1. El cuarto donde se realiza la toma debe ser de mínimo 3m x 4m.
2. El tiempo de aclimatación (con bata) de la paciente debe ser de mínimo 10 minutos y máximo 15 minutos.
3. El tiempo de aclimatación puede ser usado para la realización de la encuesta y firma de consentimiento.
4. El aire acondicionado debe encontrarse entre 18ºC y 23ºC.
5. El aire acondicionado debe encenderse al menos 1 hora antes de comenzar la toma de las imágenes.
6. Durante el registro, la paciente debe estar sentada en una silla estática y sin espaldar.
7. Durante la toma del registro, la paciente no debe tener contacto con ninguna superficie de la parte superior hacia arriba.
8. La distancia de la paciente al lente de la cámara debe ser entre 60 y 90 cm (para el caso de la cámara y lente que se tiene en la investigación).

9. Se realizarán 5 tomas termográficas de las mamas de la paciente, sin bata (como muestra la Figura 21):

    - De frente: La paciente debe colocar las manos en la cabeza y la línea de visión de la paciente debe coincidir con la línea de visión de la cámara (en realidad están a 180º)
    - Lateral derecha: la paciente debe subir los brazos y ubicarse de tal forma que su línea de visión gire noventa grados en dirección contraria a las manecillas del reloj con respecto a la línea original de frente.
    - Oblicua derecha: la paciente debe subir los brazos y ubicarse de tal forma que su línea de visión gire a un ángulo menor que noventa grados en dirección contraria a las manecillas del reloj.
    - Lateral izquierda: la paciente debe subir los brazos y ubicarse de tal forma que su línea de visión gire noventa grados en dirección de las manecillas del reloj con respecto a la línea original de frente.
    - Oblicua izquierda: la paciente debe subir los brazos y ubicarse de tal forma que su línea de visión gire a un ángulo menor que noventa grados en dirección de las manecillas del reloj.

El ángulo de las vistas oblicuas se obtiene bajo el límite máximo en el cual los senos no se superponen.

10. El registro de las 5 imágenes no puede tardar más de 15 minutos.
11. Se deben registrar la nomenclatura de las imágenes, que genera la cámara en forma automática, en el documento de cada formulario.
12. Se debe verificar el almacenamiento de 5 imágenes por paciente.
13. Al finalizar la jornada de toma de registros se deben descargar las imágenes y reinicializar la nomenclatura automática de archivos de la cámara.

3.3 RECOPILACIÓN DE LOS FACTORES SOCIO-DEMOGRÁFICOS, HEREDITARIOS, HORMONALES Y CLÍNICOS

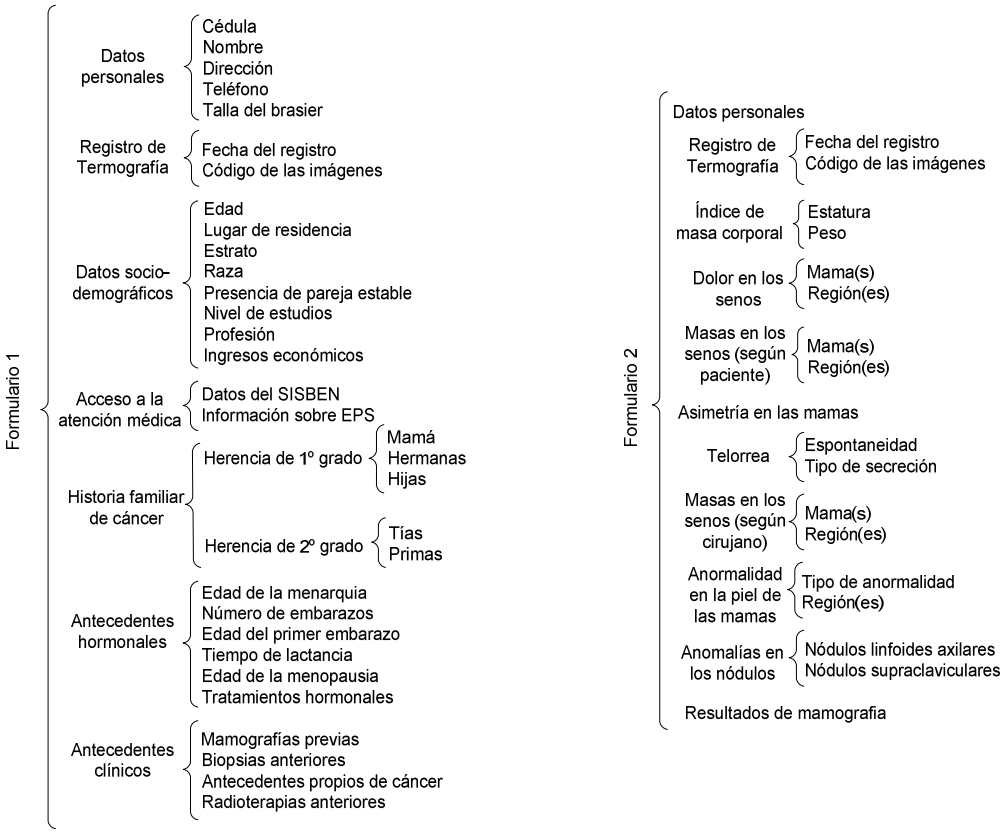
La información referente a los factores de riesgo, que pueden incidir en el desarrollo del cáncer de seno, es fundamental para revelar un posible pronóstico o diagnóstico de la enfermedad. En este trabajo se han elaborado dos formularios: el primero es una encuesta que involucra la recopilación de algunos datos socio-demográficos, hereditarios respecto al cáncer, hormonales, entre otros. El segundo reúne información semejante a una historia clínica: presencia de masas, dolor, anomalías en la piel, etc. Ambos fueron creados con la asesoría epidemiológica y clínica involucrada en el proyecto de COLCIENCIAS, teniendo en cuenta diferentes estudios sobre los factores de riesgo del carcinoma de glándula mamaría, como señala la Tabla 1. La Figura 22 muestra los esquemas resumidos de los dos formularios, mientras que en el ANEXO B se encuentra una copia completa de cada encuesta.

La información obtenida a través de los formularios, fue digitalizada mediante el software libre *EpiData Entry* [17], el cual dispone de diversas herramientas para la gestión y documentación de datos epidemiológicos. Este programa permite la verificación tras doble entrada de datos, listado de números de identificación en varios archivos, resumen de codificación, copias de seguridad con fecha, etc. Con *Epidata* se exportaron los datos a formato de texto y luego fueron retomados en MATLAB®, mediante la ejecución de un script, desarrollado en este trabajo.

En la presente investigación, se recopila información variada, que muy posiblemente no sea utilizada a cabalidad en el desarrollo del sistema de clasificación. Sin embargo, todos los datos se almacenan, para ser acumulados en el transcurso del tiempo, lo que permitirá a futuros estudios involucrar un mayor número de variables.

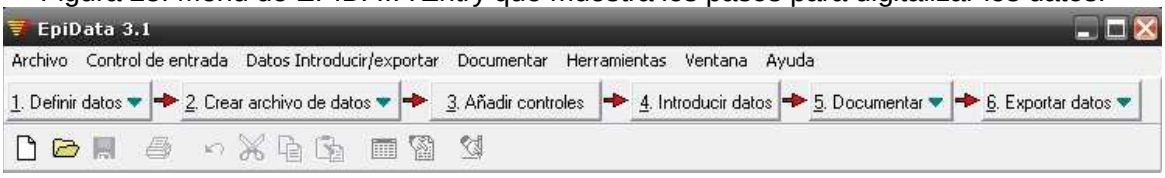
[17] Disponible en www.epidata.dk

Figura 22. Esquemas de los formularios que recopilan la información socio-demográfica, hereditaria, hormonal (Formulario 1) y fisiológica (Formulario 2).



Fuente: el autor

Figura 23. Menú de EPIDATA Entry que muestra los pasos para digitalizar los datos.



Fuente: el autor

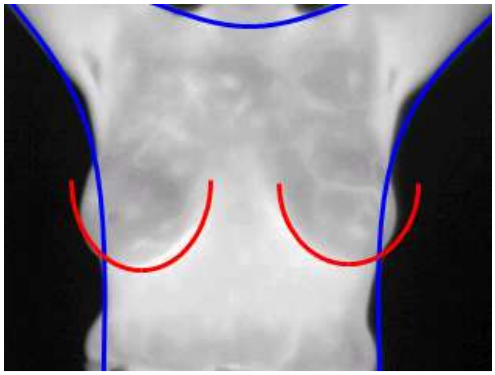
4 SEGMENTACIÓN DE TERMOGRAFIA DE MAMA MEDIANTE LA TRANSFORMADA DE HOUGH

Desde la década de 1970 se han llevado a cabo numerosos estudios en Norteamérica que han demostrado las virtudes de la termografía infrarroja, en el ámbito de la medicina, como una técnica de diagnóstico alternativa para la detección del carcinoma de glándula mamaria [8]. Sin embargo, con frecuencia, el diagnóstico que se emite empleando esta técnica se realiza con base procedimientos manuales y subjetivos por parte del personal médico [63][52].

Como es de esperarse, cualquier diagnóstico realizado de esta manera presenta una variabilidad de resultados, en función del personal que evalúe las imágenes termográficas (o termogramas), al no ser un proceso riguroso y metódico. Por tanto, una estrategia para solucionar y superar este inconveniente es someter los termogramas registrados a técnicas propias del procesamiento digital de imágenes. De este modo, el procedimiento de diagnóstico puede separarse en dos sub-procesos: uno enfocado en segmentar autónomamente las glándulas mamarias, y otro, en el cual se realizan los análisis necesarios para emitir el diagnóstico en cuestión [12][13][64][65]. En consecuencia, los estudios realizados hasta la fecha en termografía médica, para detectar esta patología en particular, pueden revestirse de rasgos objetivos como resultado de los procesos computacionales autónomos convencionalmente empleados en el tratamiento digital de imágenes.

Ahora bien, en lo que se refiere a este trabajo, se ha propuesto una metodología enfocada en solucionar el primer sub-proceso descrito líneas arriba. Para esto, se parte de un proceso preliminar de tipificación cuyo objetivo es establecer lineamientos de trabajo y/o procesamiento. Luego, se presenta una etapa intermedia de pre-procesamiento que establece condiciones propicias para la subsiguiente

Figura 24. Tipificación geométrica de la anatomía de la mama



Fuente: el autor

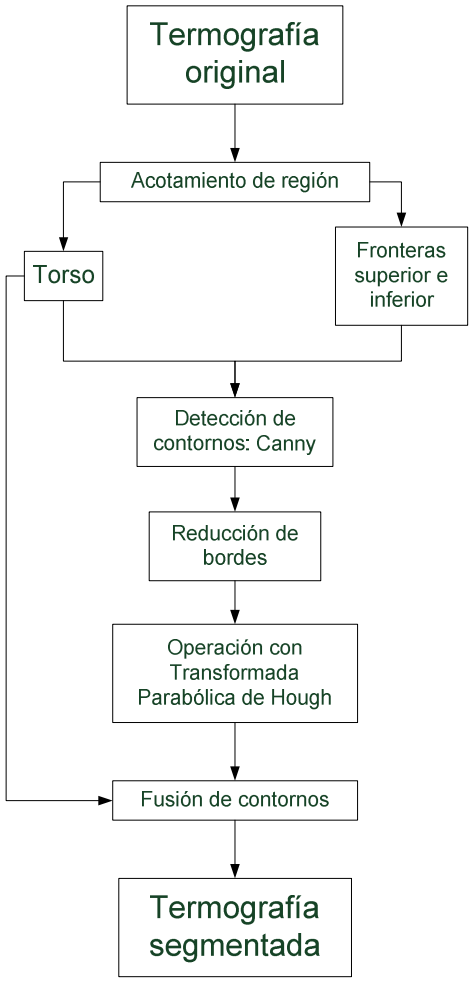
segmentación. Así pues, se entra al eje central de la metodología planteada, donde la articulación de los lineamientos sugeridos en la tipificación, junto con los resultados del pre-procesamiento, permite llevar a término una segmentación satisfactoria de las glándulas mamarias.

4.1 TIPIFICACIÓN

Como medida preliminar antes de iniciar la implementación y la ejecución de la segmentación propuesta, se ha llevado a cabo una identificación de las características de la anatomía registrada en los termogramas, para así identificar cuáles de ellas son relevantes en la resolución del problema de segmentación.

De esta manera, ha tenido lugar un proceso de tipificación que conceptualiza un modelo de lo registrado y en él, como se aprecia en la Figura 24, se han caracterizado los rasgos anatómicos mediante una representación geométrica, la cual, se centra en dos elementos primordiales: el torso y las glándulas mamarias. Estos se han modelado geométricamente a través de segmentos curvos y se muestran en distinto color para diferenciarlos. En primer lugar, representado y delimitado por segmentos de color azul, se encuentra el torso femenino. Si bien, no es la totalidad de esta área lo relevante para este trabajo sino una subregión de ella, a partir de su acotamiento

Figura 25. Metodología para la segmentación de los senos



Fuente: el autor

puede establecerse una región de trabajo más pequeña que la imagen original, que contiene las glándulas mamarias. En segundo lugar, para completar se encuentran las glándulas mamarias que corresponden al objeto de interés de este trabajo. Aquí se han representado geométricamente por medio de dos segmentos parabólicos separados, pero de construcción similar y de color rojo. A diferencia del resto del torso, en una imagen termográfica, son las glándulas mamarias quienes proveen la información acerca de la condición fisiológica estudiada. Por consiguiente, es imperativa su acertada identificación y

segmentación a partir de su representación geométrica.

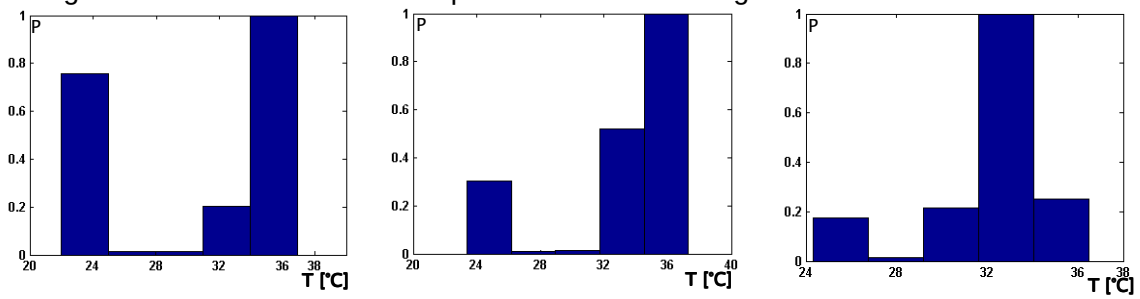
4.2 METODOLOGÍA PARA LA SEGMENTACIÓN DE MAMAS

El planteamiento propuesto ha sido diseñado con base en las observaciones logradas a partir de la tipificación mencionada. Así mismo, se hace evidente un proceso empírico orientado al tratamiento digital de la imagen térmica que se ha dividido en dos etapas: pre-procesamiento del registro térmico y segmentación del objeto de interés. En resumen, los procedimientos implementados abarcan técnicas de filtrado espacial, de binarización, de detección de bordes y la transformada de Hough para la localización de parábolas. Como tal, la estructura de la metodología propuesta puede observarse en la Figura 25.

4.2.1 Pre-procesamiento del registro térmico

En esencia, se ha formulado esta etapa con el objetivo de proveer las condiciones necesarias para asegurar el éxito de la subsiguiente etapa de segmentación. Es así como, para la consecución de esta meta y luego de una serie de experimentos, fueron identificados dos elementos para ser articulados con el proceso de segmentación. Particularmente, estos elementos son el resultado de aplicar dos procedimientos distintos, y por separado, sobre la imagen termográfica original: el acotamiento espacial de la región de trabajo y el mejoramiento del termograma mediante filtrado espacial. Es decir, el primero corresponde a una máscara binaria que define la región de trabajo acotada, y el segundo, a una versión filtrada espacialmente de la imagen térmica original. Este último procedimiento, se encuentra implícito en la detección de bordes de Canny, el cual aplica un filtro gaussiano bidimensional con una determinada desviación estándar, sobre la termografía.

Figura 26. Distribuciones de temperatura de tres termogramas de mama diferentes



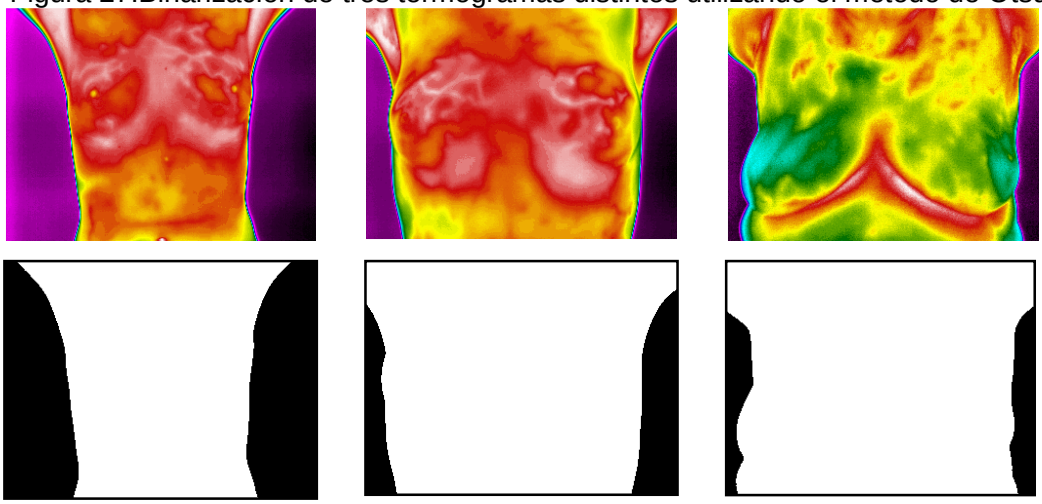
Fuente: el autor

4.2.1.1 Acotamiento de la región de trabajo

Con antelación, se ha mencionado la importancia de confinar las glándulas mamarias a una región mucho más pequeña comparada con el tamaño original de la imagen térmica. Razón por lo cual, inicialmente se ha propuesto enmarcar los senos en el área definida por la periferia del torso como se expuso en la sección tipificación. En particular, una forma eficiente de construir esta región ha sido lograda al segmentarla por completo de la escena original empleando un proceso de binarización. Se determina un umbral global, por medio del método de Otsu y luego se realiza la *umbralización,* resultando en una división de la imagen en el cuerpo de la paciente y el fondo, a través de una máscara binaria.

Este procedimiento se ha considerado viable, ya que al graficar la distribución normalizada de temperaturas de un termograma frontal de los senos, se observan dos agrupaciones claramente separadas: a la izquierda, la correspondiente al fondo de la escena, y a la derecha, la correspondiente al torso, con una temperatura media elevada comparada con la primera agrupación. Como muestra de este hecho, en la Figura 26 se exhiben las distribuciones normalizadas de temperatura que presentan tres distintos termogramas de glándulas mamarias. La Figura 27 muestra los efectos de la binarización con el método escogido, una vez fue aplicada sobre los termogramas correspondientes a las distribuciones de temperatura de la Figura 26.

Figura 27.Binarización de tres termogramas distintos utilizando el método de Otsu



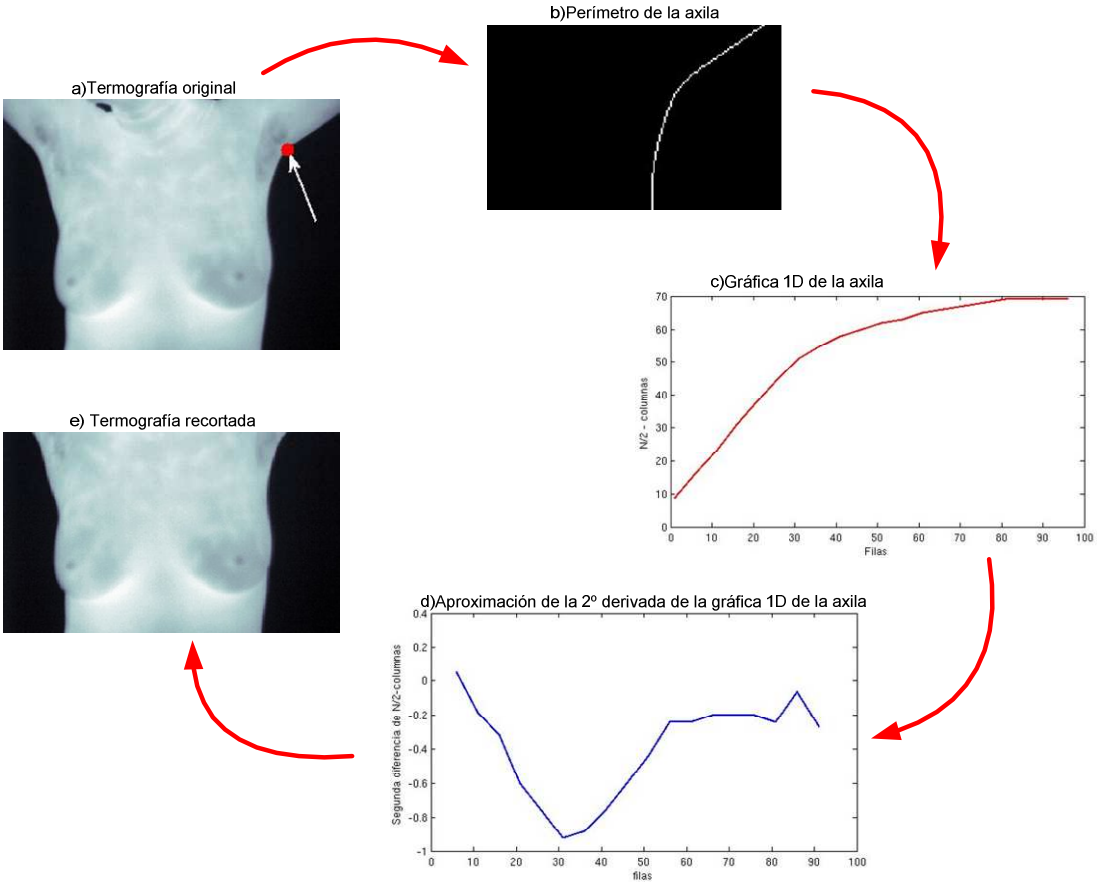
Fuente: el autor

Si bien este procedimiento ha conducido a la identificación preliminar del torso, representado por una máscara binaria, el establecimiento, además, de un corte superior y un corte inferior, acerca de la ubicación vertical de las glándulas mamarias en la imagen registrada (y por extensión en la máscara binaria obtenida) ha hecho posible acotar aun más el tamaño de la región en cuestión. El corte inferior se determina como el último veinteavo de la imagen, mientras el extremo superior se calcula con base en un algoritmo desarrollado para detectar las axilas.

4.2.1.1.1 Detección de las axilas

Si se observa el torso de la mujer en cualquier termografía de seno, es evidente que en las axilas se encuentra una curvatura bien pronunciada. Aún más, el punto medio de las axilas es de inflexión (ver Figura 28a), el cual puede ser encontrado fácilmente utilizando algunas aproximaciones del cálculo diferencial. La Figura 28 muestra todo el proceso para detectar la axila izquierda en una termografía de mama. A partir de la máscara binaria, producto de la umbralización anterior, se halla el perímetro del torso y se recorta la imagen a la región superior derecha, como expone la Figura 28b.

Figura 28. Detección de la axila del seno izquierdo en una termografía de mama

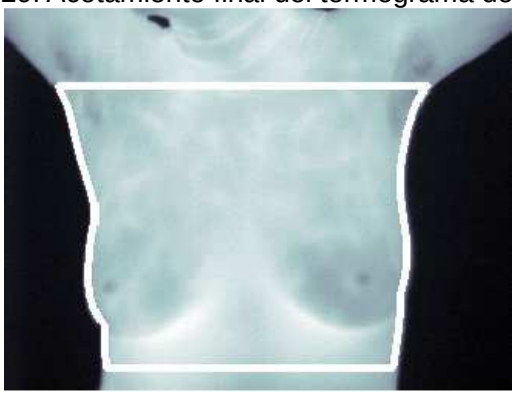


Fuente: el autor

Luego, se resta a la mitad del número de columnas de la termografía original, el valor de cada columna, obteniéndose la función unidimensional de la Figura 28c. Para determinar el punto de inflexión es necesario calcular la segunda derivada de la representación unidimensional de la axila (ver Figura 28d). Esta se aproxima con una operación de doble diferencias. Puesto que se trabaja con una función discreta, la inflexión está bien descrita por el mínimo de la función. Por último, se procede a recortar la imagen como se observa en la Figura 28e.

La Figura 29 muestra la región de trabajo establecida para el termograma anteriormente expuesto en la Figura 28 según los lineamientos descritos.

Figura 29. Acotamiento final del termograma de mama



Fuente: el autor

Para este ejemplo, se ha superpuesto el contorno (en color blanco) de la región acotada sobre el termograma en cuestión con propósitos ilustrativos.

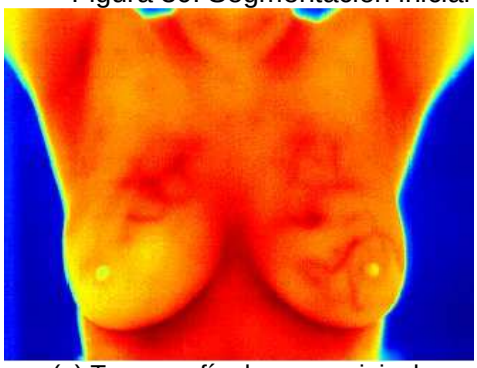
4.2.2 Segmentación de las Glándulas Mamarias

En lo que respecta a este trabajo, la solución implementada para el problema de la segmentación ha girado en torno al elemento restante de la tipificación, es decir, la caracterización de cada seno mediante una curva parabólica. Por tanto, el proceso de segmentación se ha centrado inicialmente en evidenciar el contorno que exhiben las glándulas mamarias en la imagen térmica, para luego establecer una relación entre este y su modelo geométrico.

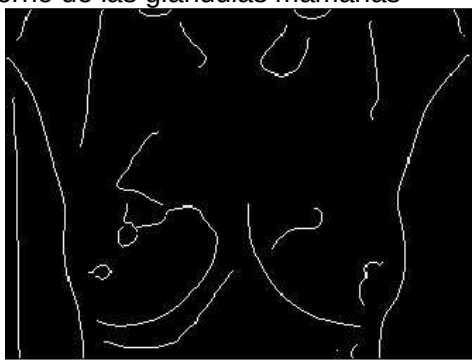
4.2.2.1 Detección de bordes

En efecto, con el objetivo de revelar el contorno de los senos en una imagen termográfica, como la de la Figura 30(a), se ha utilizado en primera instancia una técnica de detección de bordes. Ya que, a lo largo de la periferia de las glándulas mamarias, éstas presentan las típicas transiciones abruptas (en valores de temperatura) que caracterizan a los bordes, su contorno puede ser detectado. Entonces, una vez se ha aplicado el detector de bordes, concretamente el operador

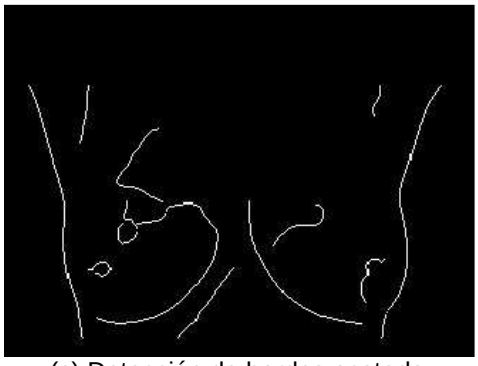
Figura 30. Segmentación inicial del contorno de las glándulas mamarias



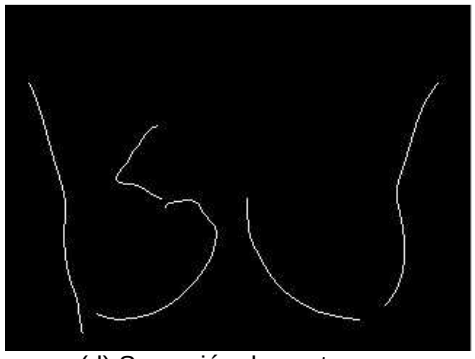
(a) Termografía de seno original



(b) Detección de bordes en la imagen total



(c) Detección de bordes acotada



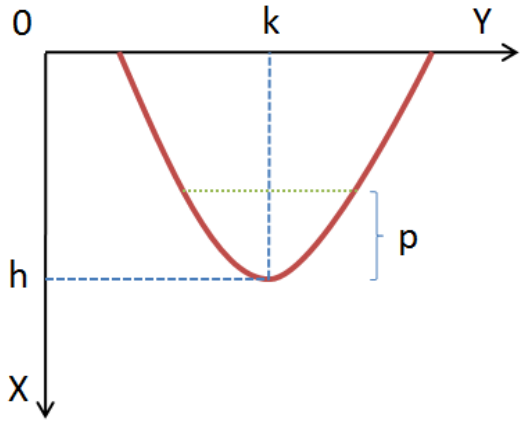
(d) Supresión de contornos

Fuente: el autor

Canny, se ha obtenido una imagen, como la de la Figura 30(b), donde se revelan parcialmente tanto el contorno de cada seno, así como, otros contornos de la anatomía registrada. Por otra parte, al articular el acotamiento con el resultado precedente se ha conseguido crear una imagen, visible en la Figura 30(c), donde los bordes detectados, que no pertenecen a la región de trabajo que fue recortada, han sido descartados.

Ahora, aunque la cantidad de contornos irrelevantes ha disminuido ostensiblemente al enfocar la atención en la región de trabajo establecida, es evidente, observando la Figura 30(c), que todavía persisten algunos en esta misma región. Pero, estos contornos remanentes, e igualmente irrelevantes, se caracterizan por su escasa longitud (en píxeles) y, por ende, en un proceso subsiguiente han podido ser suprimidos con base en ese criterio.

Figura 31. Representación cartesiana de un segmento parabólico típico, para ser detectado mediante la Transformada de Hough



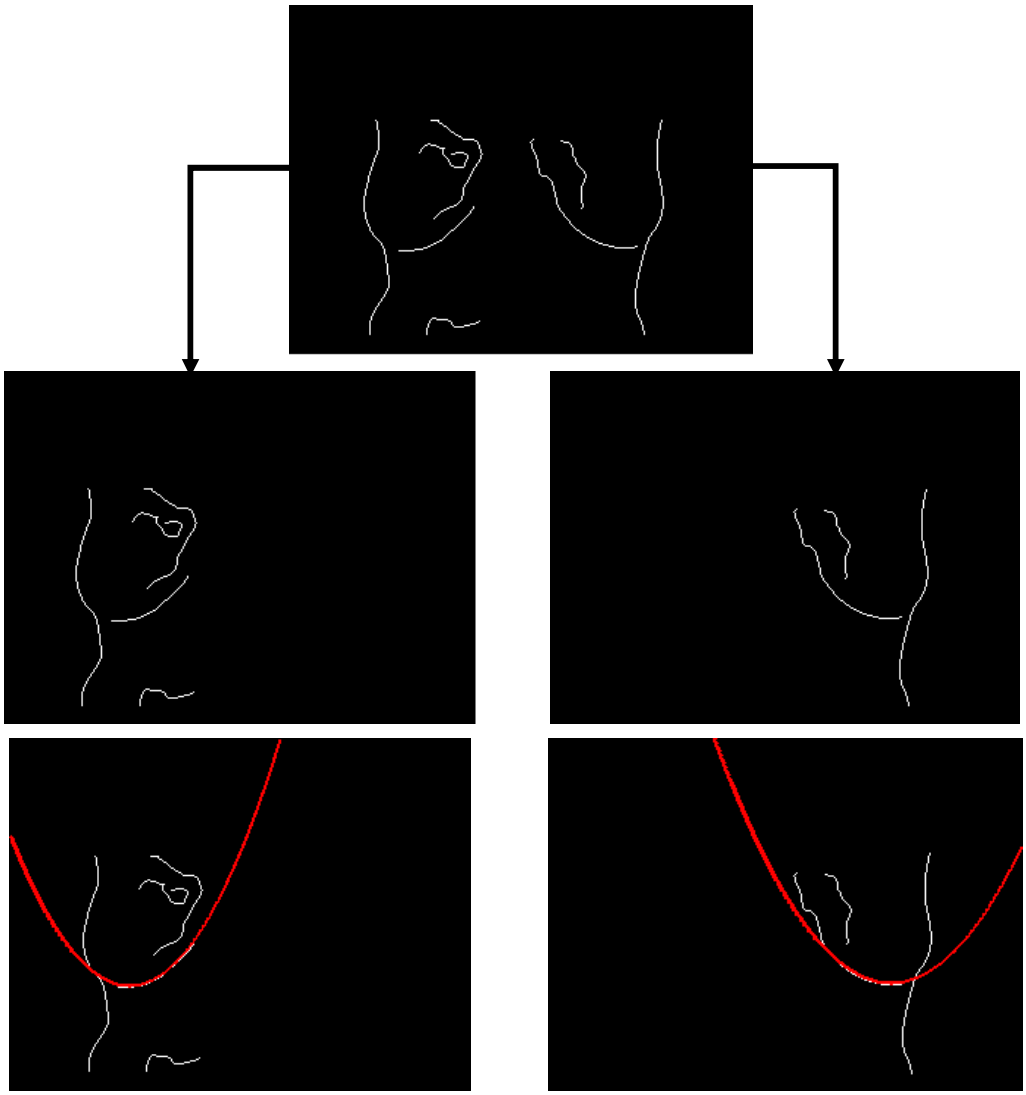
Fuente: el autor

La Figura 30(d) exhibe el resultado obtenido luego de una supresión de contornos, que ha prescindido de aquellos que no superan el 30% de la longitud del contorno más extenso presente en la imagen. Se ha obtenido una imagen en la cual sólo persisten contornos que son de interés para el proceso de segmentación, encontrándose entre ellos los que delimitan a las glándulas mamarias. Para este momento, la silueta de cada seno se ha hecho distinguible a simple vista, así como también, resulta evidente el porqué de la elección de una parábola para haber representado cada seno en la tipificación. Sin embargo, tal selección no ha de limitarse a un propósito puramente visual o representativo. En consecuencia, la forma parabólica atribuida, para efectos de este trabajo, ha de orientarse a la detección paramétrica del contorno de cada seno utilizando una versión modificada de la Transformada de Hough.

4.2.2.2 Transformada Parabólica de Hough (TPH)

La transformada de Hough es una técnica que ha sido convencionalmente utilizada para detectar líneas o curvas parametrizables en imágenes de bordes [66]. En particular, aquí se ha enfocado esta técnica para la detección de

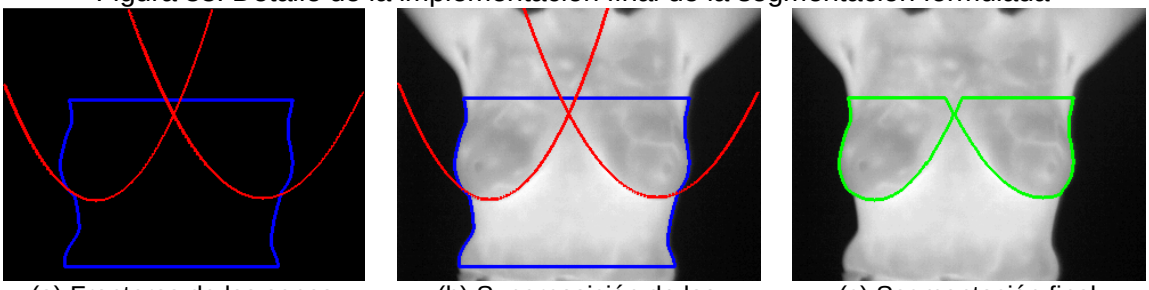
Figura 32. Proceso paso a paso de la detección con base en la Transformada de Hough



Fuente: el autor

segmentos parabólicos, específicamente aquellos que coinciden con la representación dispuesta en la Figura 31. De donde, se aprecian tres parámetros que rigen su construcción: *h* y *k* que determinan el vértice de la parábola, y *p* que determina la distancia del vértice al foco de la parábola, o la llamada distancia focal.

Figura 33. Detalle de la implementación final de la segmentación formulada



(a) Fronteras de los senos (b) Superposición de las fronteras con la imagen (c) Segmentación final

Fuente: el autor

Igualmente, en el marco de esta transformación se han involucrado las ecuaciones:

$$h = x + pt^2 \qquad (3.1)$$

$$k = y - 2pt \qquad (3.2)$$

Que constituyen la representación paramétrica de un segmento parabólico, además, del núcleo interno de la TPH.

Ahora, puesto que el objetivo de la TPH no es detectar un solo contorno parabólico, si no dos, se ha implementado un proceso paralelo de detección con miras de garantizar su éxito. Para esto, a partir de la imagen de bordes disponible se crean sendas imágenes nuevas que contienen los contornos asociados a las glándulas mamarias por separado. Acto seguido, se ha efectuado la TPH simultáneamente en cada imagen para identificar el segmento parabólico que se ajusta al contorno de cada seno. A saber, la Figura 32 resume paso a paso el proceso de detección de la forma parabólica de cada una de las glándulas mamarias.

4.2.2.3 Implementación final de la segmentación

Una vez la caracterización de cada seno ha concluido satisfactoriamente, los

respectivos modelamientos parabólicos de sus contornos pueden fusionarse entre sí, para unirse a continuación con el borde constituido sobre el tronco en la etapa de pre-procesamiento.

Llevando lo anterior a la práctica, se obtiene la superposición de fronteras expuesta en la Figura 33(a), que resulta ser más diciente si se superpone con el termograma que hasta ahora ha estado bajo análisis; visto a su vez en la Figura 33(b). Así pues, ciertamente, cada seno se ha confinado a una región óptima que resulta de la intersección de las áreas delimitadas por los distintos contornos detectados. Finalmente, en la Figura 33(c), se exhibe la segmentación final lograda mediante la metodología propuesta en este trabajo.

5 SELECCIÓN DE LAS VARIABLES

Las variables clínicas como térmicas, se han recopilado a través de las termografías de mama y los formularios que las pacientes llenan durante el proceso de registro. En secciones anteriores se han señalado los posibles factores de riesgo, así como los descriptores que pueden calcularse para cuantificar asimetrías térmicas. Sin embargo, es muy importante realizar una recopilación de todas las variables, preseleccionarlas y posteriormente analizarlas, para disminuir su número a las más relevantes, según la correlación estadística respecto al carcinoma de glándula mamaria. Este paso es fundamental para optimizar el proceso de reconocimiento, haciendo que los algoritmos implementados sean más fáciles de diseñar, más rápidos, más coherentes, y con mayor interpretación.

5.1 PRESELECCIÓN DE LAS VARIABLES

Antes de comenzar con el análisis estadístico, es primordial observar todas las posibles variables a utilizar. Puesto que, algunas de estas, pueden ser información útil o por el contrario no representan alguna descripción significativa, a primera vista. Este procedimiento se divide de acuerdo al tipo de variable: los parámetros térmicos poseen un carácter preciso o numérico, mientras los factores clínicos tienen una representación más nominal o categórica.

5.1.1 Preselección de las variables termográficas

El tratamiento de las variables termográficas es más sencillo, dado que se definen matemáticamente a partir de los pixeles de una *región térmica* de la imagen original. Dichas regiones, son obtenidas a partir de las máscaras binarias derivadas del proceso de segmentación. En la sección 1.3.6 se mostró una cantidad apreciable de posibles descriptores para detallar el comportamiento térmico de los senos. Sin embargo, no

Tabla 6. Variables térmicas preseleccionadas

Descriptores de 1° orden	Descriptores de 2° orden
Media	Energía
Moda	Contraste
Mediana	Correlación
Desviación estándar	Varianza
Asimetría o Sesgo	Entropía
Curtosis	Varianza de la diferencia
Energía	Homogeneidad
Entropía	
Máximo	
Rango	

todos los descriptores de segundo orden fueron escogidos, con el fin de mostrar un mayor análisis de *asimetrías* en el patrón térmico de un seno respecto al otro, como lo detallan varios estudios para detección de cáncer de seno con termografía infrarroja [67][68]. Los descriptores termográficos preseleccionados se muestran en la Tabla 6.

Tabla 7. Variables clínicas preseleccionadas

Variable	**Forma de medición**
Edad	Se mide en años
Menarquia	Se mide en años
Edad del primer embarazo	Binaria: 1 si el primer niño nació después de los 30 años o en caso de nuliparidad; 0 en el resto [69].
Edad Menopausia	Binaria: 0 si se presento la menopausia antes de los 50 años o si aún no se ha presentado; 1 en el resto [69].
Índice de masa corporal	Peso/Estatura² en kg/m²
Dolor en los senos	Binaria: 1 si hay Dolor; 0 si no.
Asimetría en las mamas	Binaria: 1 si existe una gran asimetría entre las mamas; 0 si no (se tolera una pequeña asimetría).
Masas en los senos (percepción del cirujano)	Binaria: 1 si hay presencia de masas mayores a 2 cm; 0 si no hay.
Anormalidad en la piel de las mamas	Binaria: 1 si hay anormalidad; 0 si no.

En general, todas las variables termográficas que serán introducidas en el modelo, corresponden al valor absoluto de la diferencia entre los descriptores de un seno respecto al otro.

5.1.2 Preselección de las variables clínicas

Las variables clínicas consisten en la información socio-demográfica, hereditaria, hormonal y fisiológica, obtenidas a través de los formularios mostrados en el ANEXO B. El tratamiento para escoger, en primera medida, estos factores, depende del tamaño de la muestra, de la veracidad de estas variables en las investigaciones epidemiológicas y del alcance de este estudio.

Las encuestas están diseñadas para recopilar más datos de los que se requiere en este trabajo, puesto que, futuros estudios epidemiológicos podrían hacer uso de esta información, en especial, cuando la muestra de pacientes aumente significativamente. Respecto a este último punto, la preselección de las variables clínicas estuvo limitada a la representación de las mismas, puesto que la cantidad de mujeres que participaron en el estudio fue de 29. Por ejemplo, la herencia de cáncer de mama de primera línea sólo se presentó en una paciente. En la Tabla 7 se pueden observar todas las variables clínicas preseleccionadas; esta señala la forma en que se mide cada variable, teniendo en cuenta diferentes investigaciones epidemiológicas (ver Tabla 1).

5.2 SELECCIÓN ESTADÍSTICA DE LAS VARIABLES

Existen muchas técnicas de *selección de variables*[18]: mediante coeficientes de

[18] Conocida en inglés como *Feature Selection*

correlación lineal (Pearson y Spearman), selección secuencial de características[19] [70], ranking a través de diferentes criterios como: distancias probabilísticas, pruebas t y área bajo la curva ROC [47], etc. Como se mencionó anteriormente, algunas variables clínicas son categóricas, por lo tanto se debe tratar de emplear métodos no paramétricos. Aunque, como las variables categóricas empleadas son dicotómicas, se pueden usar algunas técnicas habituales sin una marcada diferencia. Más importante, es el hecho de que la variable dependiente "el resultado histopatológico" es binaria (hay cáncer o no hay cáncer).

5.2.1 Correlación estadística mediante los coeficientes de Pearson y Spearman

Como medida preliminar, se determinaron los coeficientes de Pearson y Spearman, para determinar las relaciones que existen, de forma independiente, entre cada variable y el resultado de la biopsia (se recomienda revisar la sección 1.5). Como *ilustración*, la Tabla 8 muestra las primeras 9 variables, ordenadas ascendentemente según el valor-p (disminuye la significancia) y de forma descendente según el coeficiente de Pearson o Spearman. Los descriptores termográficos de segundo orden, fueron obtenidos mediante la matriz de co-ocurrencia en dirección horizontal (este). Se observa que la *Edad* y la *Menopausia* presentan una marcada correlación, mientras la *Media* y la *Homogeneidad* tienen una correlación aceptable respecto a la presencia o ausencia de cáncer de mama.

5.2.2 Ranking de las variables a través del área bajo la curva ROC (AUC[20])

La curva ROC es una herramienta que se usa en la verificación de sistemas de clasificación, validación de modelos epidemiológicos y como prueba estadística no paramétrica. El área bajo la curva ROC varia entre *1* (separación completa de las

[19] Stepwise regression, en inglés

[20] En inglés *area under the ROC curve*

funciones de distribución de probabilidad de cada clase) y ½ (superposición máxima de las funciones de distribución de probabilidad). De esta forma, la AUC es una medida de la capacidad de discriminación entre clases de una variable específica. Experimentalmente, la curva ROC puede ser construida realizando un barrido de umbrales y calculando los porcentajes de clasificaciones correctas y erróneas sobre los vectores que almacenan la variable [47].

En este trabajo, se usa la AUC como parámetro para realizar un ranking de las variables. No obstante, este es un criterio *independiente* de separación entre clases, es decir, que se aplica entre cada variable independiente y la variable dependiente, de forma separada. Para tener en cuenta las relaciones que puedan existir entre características, se introduce un parámetro α que usa la información de correlación para ponderar el área bajo la curva ROC de cada variable, mediante el factor:

$$AUC\,[1 - \alpha\,\rho]$$

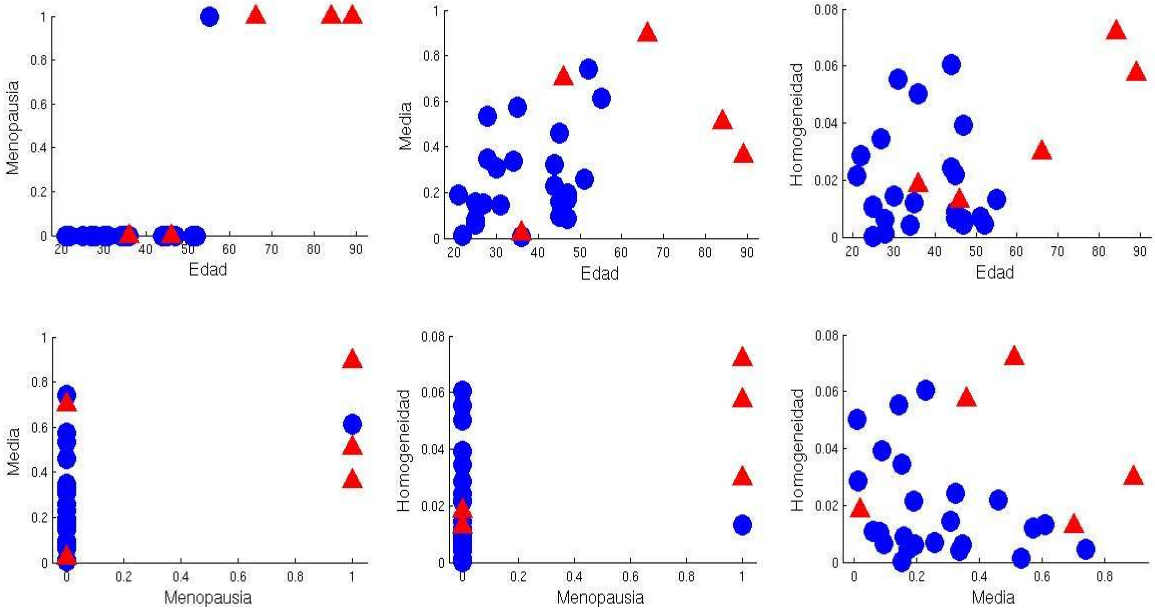
Donde ρ es el promedio de la magnitud de los coeficientes de correlación cruzada entre el candidato y todas las variables anteriormente seleccionadas, α es un valor escalar entre *0* y *1*, cuando es *0* las características no se ponderan. Un valor grande de ρ (cercano a 1) realza la significancia estadística; es decir, que las variables que están altamente correlacionadas con las previamente elegidas, son menos probables de ser incluidas en la lista de salida [47][71].

Tabla 8. Ilustración de la correlación lineal entre las variables y el resultado histopatológico

Variable	**Pearson**	**Valor – p**
Edad	0,6256	0,0003
Menopausia	0,6116	0,0004
Media	0,3851	0,0391
Homogeneidad	0,3732	0,0461
IMC	0,2781	0,1441
Curtosis	-0,2766	0,1463
Alteraciones piel	0,2750	0,1488
Máximo	0,2551	0,1816
Menarquia	0,2281	0,2339

Variable	**Spearman**	**Valor – p**
Menopausia	0,6116	0,0004
Edad	0,4646	0,0111
Homogeneidad	0,3710	0,0476
Media	0,3055	0,1070
Alteraciones piel	0,2750	0,1488
Curtosis	-0,2728	0,1523
Menarquia	0,2701	0,1565
IMC	0,2619	0,1700
Máximo	0,2215	0,2482

Figura 34. Diagramas de dispersión de algunas variables con alta correlación



Fuente: el autor

En la Tabla 9 se muestra un ejemplo de utilización del ranking mediante el área bajo la curva ROC (en el capítulo 6 se muestra el análisis sobre todo el conjunto de datos), teniendo en cuenta la correlación entre variables ($\alpha = 0.6$). El orden que muestra la tabla es en forma decreciente según la *AUC ponderada*, pero se recalca el valor real del área. Se aplicó este método no paramétrico sobre todas las variables termográficas y clínicas, y en ambos casos por separado.

De nuevo, se puede observar que las primeras variables, en su mayoría, corresponden a las descritas por el coeficiente de Pearson (o Spearman). A pesar de, que en esta ocasión se utilizó la matriz de co-ocurrencia en dirección vertical norte. Para investigar algo más su comportamiento, se graficaron los diagramas de dispersión de las primeras variables en la Figura 34. Aún con las limitaciones de una muestra pequeña, se nota una buena discriminación entre las pacientes sanas y enfermas, sobre todo a través de la edad y la menopausia. Estas dos pueden separar las clases por medio de un umbral directo. Sin embargo, en casi

Tabla 9. Ranking según la AUC para: todas las variables, sólo las termográficas y únicamente las clínicas. La dirección de la matriz de co-ocurrencia es vertical.

Todas las variables	
Variable	**AUC**
Edad	0,854
Homogeneidad	0,842
Menopausia	0,779
Curtosis	0,708
Media	0,733
Desviación estándar	0,667
Alteraciones piel	0,638
Menarquia	0,700
IMC	0,700
Máximo	0,667
Varianza(dif)	0,617
Mediana	0,629
Energía(1ºorden)	0,608
Rango	0,617
Moda	0,583
Contraste	0,600
Sesgo	0,592
Entropía(1ºorden)	0,592
Energía(2ºorden)	0,558
Entropía(2ºorden)	0,558
Varianza(2ºorden)	0,542
Asimetría	0,496
Correlación	0,508
Primer embarazo	0,433
Dolor	0,450
Masas	0,463

Termográficas	
Variable	**AUC**
Homogeneidad	0,842
Curtosis	0,708
Media	0,733
Desviación estándar	0,667
Máximo	0,667
Varianza(dif)	0,617
Moda	0,583
Mediana	0,629
Contraste	0,600
Energía(1ºorden)	0,608
Rango	0,617
Sesgo	0,592
Energía(2ºorden)	0,558
Entropía(1ºorden)	0,592
Varianza(2ºorden)	0,542
Correlación	0,508
Entropía(2ºorden)	0,558

Clínicas	
Variable	**AUC**
Edad	0,854
Menopausia	0,779
Menarquia	0,700
Alteraciones piel	0,638
IMC	0,700
Asimetría	0,496
Primer embarazo	0,433
Dolor	0,450
Masas	0,463

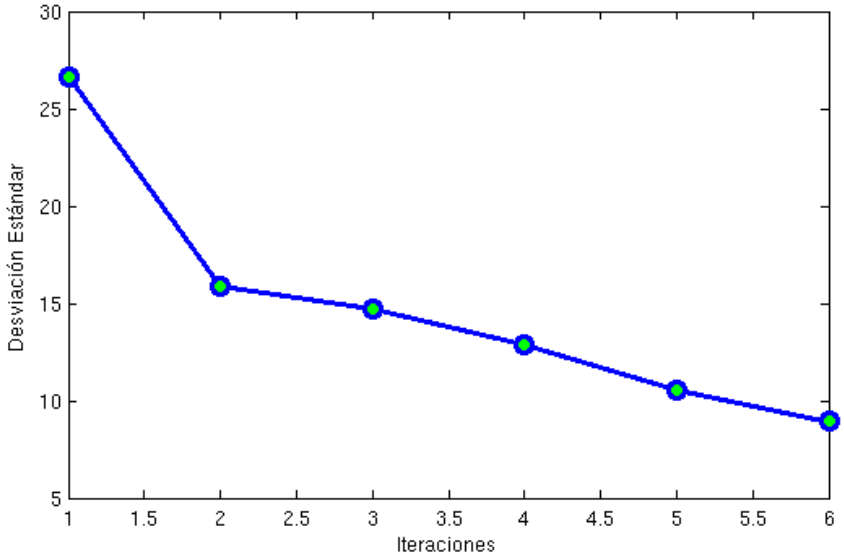
todas las comparaciones dos resultados con carcinoma se *mezclan* junto a los demás resultados benignos.

5.2.3 Selección secuencial de variables

Este método utiliza un algoritmo de búsqueda, hacia delante o hacia atrás. Esta técnica iterativamente añade o remueve las variables, hasta que algún criterio de finalización se cumple. Por ejemplo, la *selección secuencial hacia delante* (SFS[21])

[21] En inglés, Sequential Forward Selection

Figura 35. Desviación estándar de cada paso en la regresión logística secuencial, aplicada a todas las variables



Fuente: el autor

y la *selección secuencial hacia atrás* (SBS[22]) son algoritmos de este tipo. Asumiendo que *D* es el número total de variables, *d* es la cantidad a seleccionar y *n* es el número de características originales. SFS es un método donde una variable, que satisface un criterio determinado, es adherida al subconjunto de variables. Esto se repite hasta que el número de éstas llegué a *d* (o algún otro criterio). SBS es una aproximación donde las variables son removidas del conjunto total, una a una, hasta que *D* – *d* variables hayan sido borradas. Tanto en SFS como en SBS, el número de subconjuntos a inspeccionar es $n + (n-1) + (n-2) +$ ... $(n-d+1)$, [70]. Se utilizó la desviación estándar de la regresión logística, como el criterio para realizar la selección secuencial hacia delante. Se usa SFS pues computacionalmente es menos costosa, pero aún más, porque la regresión logística de muchos parámetros con una cantidad de muestra pequeña, no converge. Como ejemplo, se tomaron las diez primeras variables del ranking general de la Tabla 9 y se limitó a *5* el número de selecciones

[22] En inglés, Sequential Backward Selection

por el método SFS, obteniéndose la inclusión de las variables: *Edad, Homogeneidad, Curtosis, Media y Alteraciones en la piel*. En la Figura 35, se describe el comportamiento de la desviación estándar de la regresión, aplicada a cada subconjunto en su respectiva iteración.

En general, cualquiera de las tres formas señaladas en este capítulo, es suficiente para determinar una selección coherente de las variables. La mayor limitación se debe al número de pacientes disponibles para el estudio. Por lo tanto, en la continuación de este proyecto, si se aumenta significativamente la población, se puede mejorar el análisis, realizando por ejemplo, validaciones y pruebas en divisiones independientes de la muestra. Esto permite una mayor precisión a la hora de escoger las variables que contienen mayor relación (y menos dependencia entre ellas) con el resultado histopatológico.

6 CONJUNTOS DE DATOS

En el capítulo 2, se detalló la forma en que se obtuvo la información, mientras en el capítulo 4 se mostraron las variables escogidas del conjunto total y un posterior análisis para la selección definitiva. Sin embargo, no se ha descrito aún, como se procedió para la generación de los datos que el sistema de inteligencia artificial utilizará para el proceso de clasificación. Por un lado, los datos clínicos que están compuestos por los factores socio-demográficos, hereditarios, hormonales y fisiológicos, ya están descritos en su totalidad, dado que la información por paciente es única y proviene de las encuestas. Por otro lado, las variables termográficas dependen de las regiones que el proceso de segmentación genera. Como muestra la Figura 36, en este trabajo se emplearon dos tipos de segmentación:

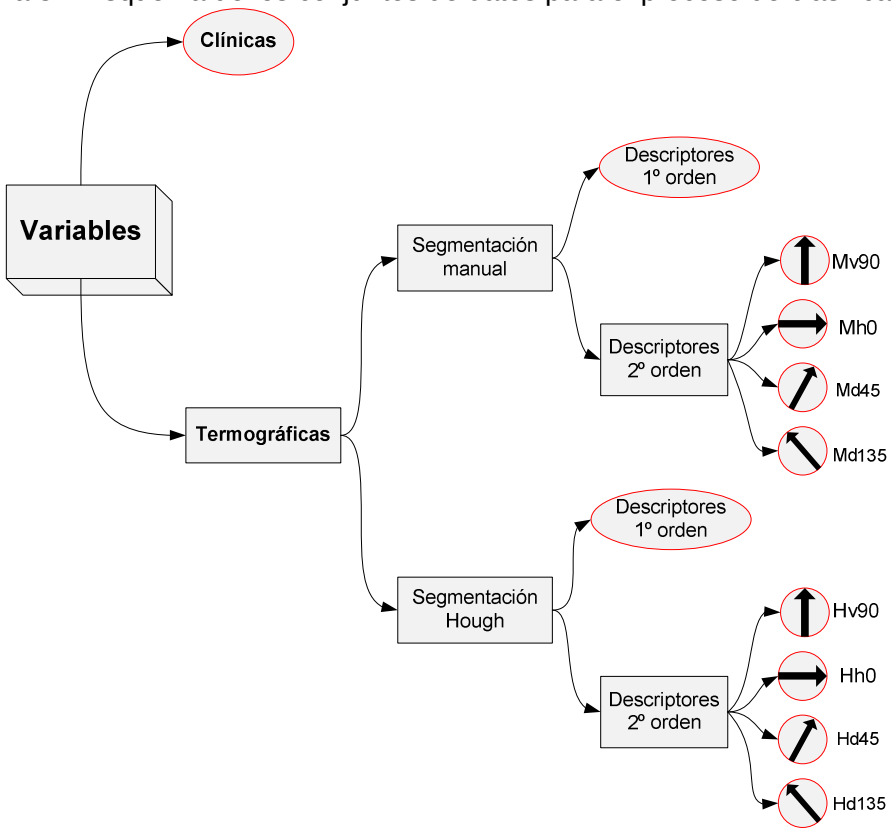
- Segmentación manual, con base en regiones elípticas realizadas por el usuario.
- Segmentación mediante la transformada de Hough parabólica y la fusión de contornos.

Figura 36. Comparación entre los dos tipos de segmentación de los termogramas de mama

(a) Segmentación manual (b) Segmentación con transformada de Hough

Fuente: el autor

Figura 37. Esquema de los conjuntos de datos para el proceso de clasificación



Fuente: el autor

A su vez, los descriptores de primer orden, por definición, se computan directamente a partir de las regiones de los senos (con la temperatura). Por otra parte, los descriptores de segundo orden dependen de la dirección elegida para el cálculo de la matriz de co-ocurrencia. Se escogieron cuatro direcciones: horizontal este, vertical norte, diagonal noreste y diagonal noroeste. Para cada una de estas se obtiene un conjunto de variables *diferentes*. La Figura 37 muestra el esquema en que se "divide" la información. Se recalca que las elipses con borde rojo, representan un grupo de variables independiente, es decir, que no se pueden subdividir o calcular en forma distinta. De acuerdo con toda la descripción anterior, para cada metodología de segmentación, se pueden formar 8 conjuntos de datos, uno por cada dirección de la matriz de co-ocurrencia. Cada uno de ellos, es un conjunto de prueba en el estudio del mejor sistema de clasificación.

Por ejemplo, un posible grupo son las variables clínicas y, los descriptores termográficos de primer y segundo orden (matriz de co-ocurrencia en dirección horizontal) derivados de las máscaras producidas mediante segmentación manual.

Puesto que, el número de pacientes que participan en la investigación es de 29 (con biopsia), se opta por el método del área bajo la curva ROC junto al factor de peso por correlación, en la selección de las variables. Además, el coeficiente de Pearson (o Spearman) no valora relaciones entre más de dos factores y la selección secuencial sería mucho más útil con una muestra mayor. Las tablas 10 y 11 ilustran el ranking de las diez primeras variables, para las regiones generadas por segmentación con transformada parabólica de Hough (incluyendo la información clínica). Se describe el orden de importancia según la AUC con ponderación $\alpha = 0.6$ (este valor no modifica abruptamente el ranking pero incluye la posible correlación entre variables), en las dos columnas que están antes de la triple línea de separación. A la derecha de esta, se encuentra el ranking exclusivo de las variables termográficas, empleando el mismo método. Asimismo, las Tabla 12 y Tabla 13 muestran los ítems para las regiones segmentadas de forma manual. Las variables clínicas fueron categorizadas anteriormente, en la Tabla 9.

Tabla 10. Ranking de las variables mediante AUC, con matriz de co-ocurrencia horizontal y vertical, para las regiones segmentadas mediante transformada de Hough

Dirección Horizontal – Este: H0			**Dirección Vertical – Norte: V90[23]**		
Todas Variables	**AUC**	**Termográficas**	**Todas Variables**	**AUC**	**Termográficas**
Edad	0,854	Homogeneidad	Edad	0,854	Homogeneidad
Menopausia	0,779	Media	Homogeneidad	0,842	Curtosis
Homogeneidad	0,783	Curtosis	Menopausia	0,779	Media
Curtosis	0,708	Desviación estándar	Curtosis	0,708	Desviación estándar
Media	0,733	Máximo	Media	0,733	Máximo
Desviación estándar	0,667	Moda	Desviación estándar	0,667	Varianza(dif)
Alteraciones piel	0,638	Contraste	Alteraciones piel	0,638	Moda
Menarquia	0,700	Mediana	Menarquia	0,700	Mediana
IMC	0,700	Energía(1ºorden)	IMC	0,700	Contraste
Máximo	0,667	Varianza(dif)	Máximo	0,667	Energía(1ºorden)

[23] V90 es vertical, 90º en dirección contraria a las manecillas del reloj del este

Tabla 11. Ranking de las variables mediante AUC con matriz de co-ocurrencia diagonal, para las regiones segmentadas mediante transformada de Houhg

Dirección Diagonal – Noreste: D45			Dirección Diagonal – Noroeste: D135		
Todas Variables	**AUC**	**Termográficas**	**Todas Variables**	**AUC**	**Termográficas**
Edad	0,854	Varianza(dif)	Edad	0,854	Homogeneidad
Menopausia	0,779	Energía(2°orden)	Menopausia	0,779	Curtosis
Varianza(dif)	0,767	Media	Curtosis	0,708	Media
Curtosis	0,708	Curtosis	Homogeneidad	0,742	Desviación estándar
Media	0,733	Contraste	Media	0,733	Máximo
Energía(2°orden)	0,692	Máximo	Desviación estándar	0,667	Moda
Alteraciones piel	0,638	Desviación estándar	Menarquia	0,700	Contraste
Menarquia	0,700	Moda	Alteraciones piel	0,638	Varianza(2°orden)
Desviación estándar	0,667	Mediana	IMC	0,700	Entropía(2°orden)
Contraste	0,675	Entropía(2°orden)	Máximo	0,667	Mediana

Tabla 12. Ranking de las variables mediante AUC, con matriz de co-ocurrencia horizontal y vertical, para las regiones segmentadas manualmente

Dirección Horizontal – Este: H0			Dirección Vertical – Norte: V90		
Todas Variables	**AUC**	**Termográficas**	**Todas Variables**	**AUC**	**Termográficas**
Edad	0,854	Curtosis	Edad	0,854	Curtosis
Menopausia	0,779	Media	Menopausia	0,779	Media
Curtosis	0,758	Energía(2°orden)	Curtosis	0,758	Energía(1°orden)
Energía(2°orden)	0,725	Moda	Energía(1°orden)	0,717	Moda
Media	0,733	Energía(1°orden)	Media	0,733	Energía(2°orden)
Energía(1°orden)	0,717	Homogeneidad	Moda	0,671	Mediana
Moda	0,671	Mediana	Alteraciones piel	0,638	Varianza(2°orden)
Alteraciones piel	0,638	Varianza(dif)	Mediana	0,696	Correlación
Mediana	0,696	Varianza(2°orden)	Menarquia	0,700	Homogeneidad
Menarquia	0,700	Contraste	Energía(2°orden)	0,658	Desviación estándar

Tabla 13. Ranking de las variables mediante AUC, con matriz de co-ocurrencia diagonal, para las regiones segmentadas manualmente

Dirección Diagonal – Noreste: D45			Dirección Diagonal – Noroeste: D135		
Todas Variables	**AUC**	**Termográficas**	**Todas Variables**	**AUC**	**Termográficas**
Edad	0,854	Curtosis	Edad	0,854	Curtosis
Menopausia	0,779	Media	Menopausia	0,779	Media
Energía(2°orden)	0,750	Energía(2°orden)	Curtosis	0, 758	Energía(1°orden)
Curtosis	0,758	Moda	Energía(1°orden)	0,717	Moda
Media	0,733	Energía(1°orden)	Media	0,733	Contraste
Energía(1°orden)	0,717	Correlación	Moda	0,671	Mediana
Moda	0,671	Entropía(2°orden)	Alteraciones piel	0,63 8	Entropía(2°orden)
Entropía(2°orden)	0,708	Mediana	Mediana	0,696	Varianza(2°orden)
Alteraciones piel	0,638	Desviación estándar	Menarquia	0,700	Homogeneidad
Correlación	0,675	Varianza(2°orden)	Contraste	0,650	Máximo

7 ANALISIS DE RESULTADOS

7.1 VALIDACIÓN DEL ALGORITMO DE SEGMENTACIÓN CON TRANSFORMADA DE HOUGH PARABÓLICA

La metodología propuesta para la segmentación automática fue puesta a prueba utilizando un conjunto de 100 imágenes termográficas que registran una perspectiva frontal de las glándulas mamarias y que fueron adquiridas siguiendo el protocolo de registro. Así mismo, estos registros corresponden a 59 mujeres que colaboraron voluntariamente en una prueba experimental y a 41 pacientes que hacen parte real de este estudio (tendrán resultado de biopsia). Las edades de todas las participantes oscilan entre los 18 y 90 años.

Ahora bien, el veredicto emitido para cada una de las segmentaciones efectuadas categoriza el resultado obtenido según la eficacia lograda, es decir: se segmenta correctamente ambas glándulas mamarias, sólo una de ellas o ninguna de ellas. Las segmentaciones fueron realizadas con una desviación estándar del filtro gaussiano de 1.8, para la detección de bordes. Sin embargo, para todas las segmentaciones parciales existe un valor alterno de desviación que permite lograr la segmentación correcta de las mamas. Por otro lado, la segmentación de las mamas fallan debido principalmente a un mal registro termográfico y/o a que los senos de la mujer presentan una forma anatómica difícil de detectar visualmente. Así pues, los resultados experimentales obtenidos se consignan en la Tabla 14.

Finalmente, en lo referente a la naturaleza de la validación cometida, esta se ha hecho de forma subjetiva puesto que no hay un método establecido para delimitar fielmente el área que superficialmente corresponde a las glándulas mamarias.

Tabla 14. Resultados experimentales de la segmentación implementada

Veredicto	Cantidad
Segmentación completa	74
Segmentación parcial	18
Segmentación fallida	9
Total	100

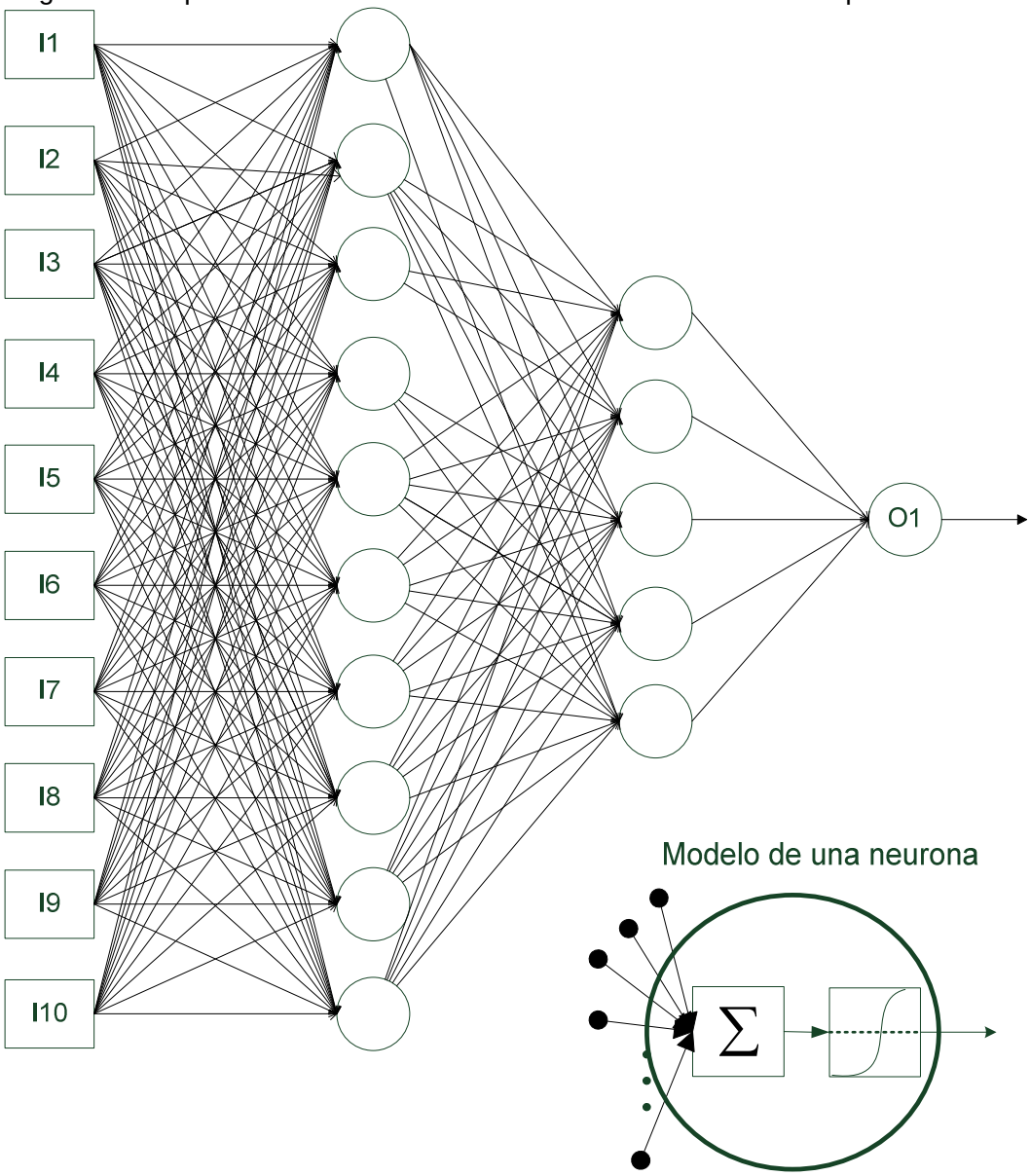
7.2 ALGORITMOS DE CLASIFICACIÓN

La Figura 14 describió las etapas de un sistema de clasificación. En este punto de la investigación, se ha logrado recopilar la información térmica y clínica, a partir del registro termográfico y las encuestas. Posteriormente, se generaron los descriptores termográficos y se fijaron los datos clínicos a utilizar, para luego seleccionar las variables más importantes referentes al resultado histopatológico de las pacientes en estudio. Ahora, se debe diseñar el clasificador, para lo cual, se implementaran varios tipos de algoritmos. Por el limitado tamaño de la muestra, el aprendizaje supervisado realizado no incluye un entrenamiento, validación y prueba, como usualmente se efectúa para evitar sobre-entrenamientos, sino que, se divide aleatoriamente el conjunto de datos en un 60% para entrenar y un 40% para validar. En el primero, siempre se incluyen tres casos con cáncer, el segundo, asiduamente contiene dos pacientes con carcinoma de mama. Este proceso se repite 15 veces, bajo el mismo algoritmo (con las mismas condiciones iníciales) y se seleccionan los resultados con mayor área bajo la curva ROC. De esta forma, se evalúa el desempeño del sistema de clasificación.

7.2.1 Red neuronal feedforward - backpropagation

Este tipo de arquitectura y algoritmo de aprendizaje es uno de los más utilizados en la literatura. En la detección de cáncer de mama con termografía infrarroja en combinación con datos clínicos, ha sido empleado por Ng, et al [52].

Figura 38. Arquitectura de la red neuronal feedforward BP utilizada para clasificar



Fuente: el autor

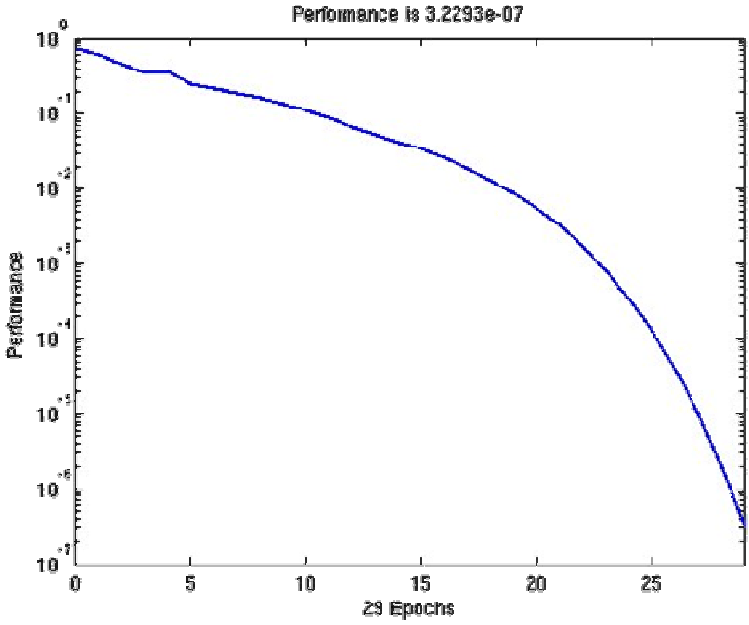
En la presente investigación se usa la configuración 10 – 10 – 5 – 1, con funciones de transferencia *tangente hiperbólica* en todas las capas, como se muestra en la Figura 38. Esta red neuronal es entrenada con el algoritmo de aprendizaje *Resilient Backpropagation* (RP), con los parámetros descritos en la Tabla 15.

Tabla 15. Parámetros usados en el algoritmo de aprendizaje Resilient Backpropagation [72]

Nombre	Valor	Significado
epochs	1000	Máximo número de iteraciones en el entrenamiento
goal	1,00E-06	Límite de la función error
min_grad	1,00E-010	Mínimo valor del gradiente
lr	0,01	Tasa de aprendizaje
delt_inc	1,2	Incremento de los pesos
delt_dec	0,5	Decremento de los pesos
delta0	0,07	Cambio inicial de los pesos
delta_max	50	Cambio máximo de los pesos

Cada grupo de datos, provenientes de los tipos de segmentación y de la dirección de la matriz de co-ocurrencia (H0, V90, D45 y D135), se utilizan para entrenar la red RP, dividiendo la muestra. La Figura 39 exhibe como converge la función de error (cuadrática) a través de un entrenamiento utilizando las variables clínicas y termográficas en dirección horizontal, provenientes de la segmentación automática. En este caso el aprendizaje terminó en la iteración 29, debido a que el error alcanzó un valor mínimo (3.23 x 10^{-7}), por debajo del umbral.

Figura 39. Función de error durante el entrenamiento de la red RP para Hh0



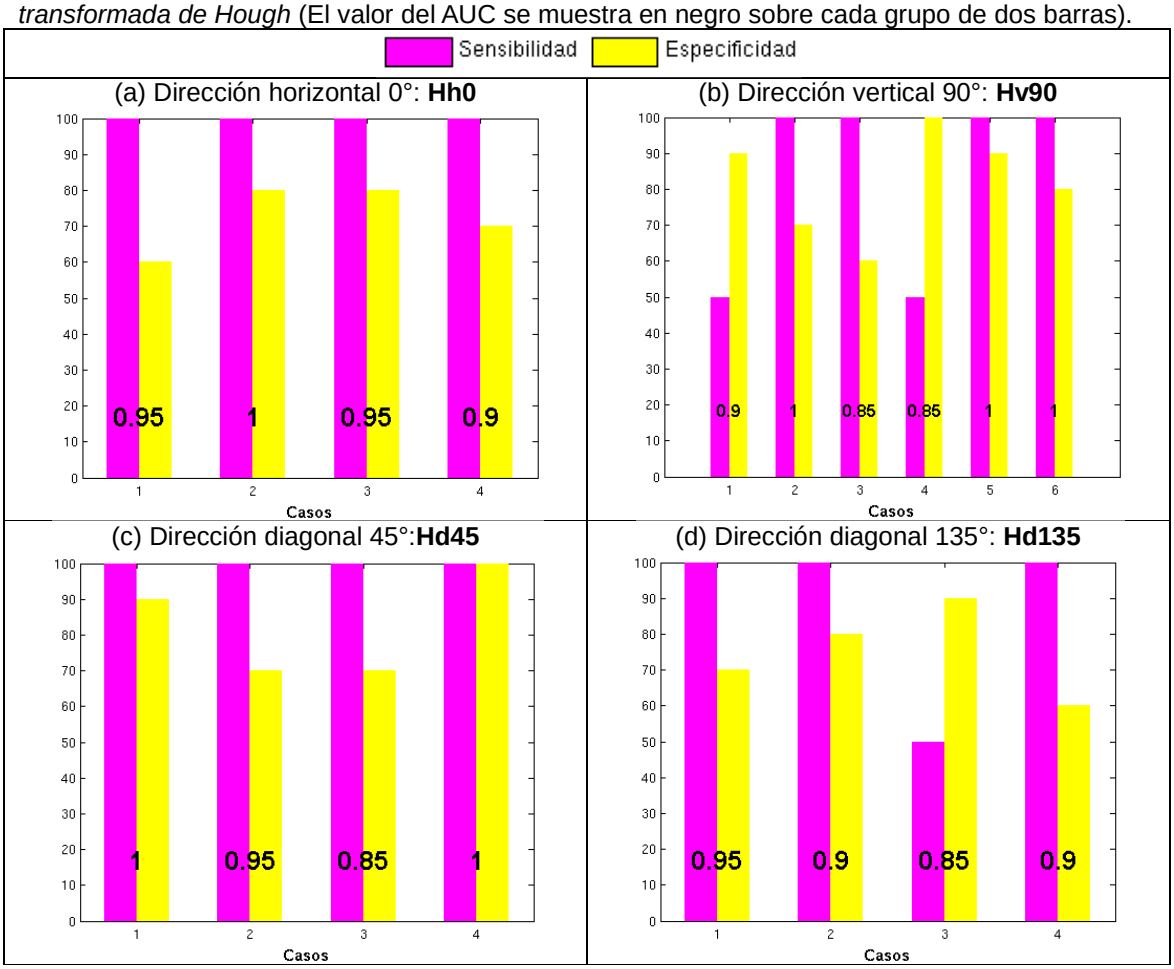
Fuente: el autor

7.2.1.1 Pruebas para la red neuronal feedfordward backpropagation con aprendizaje RP

Como anteriormente se comentó, cada conjunto de prueba fue dividido de forma aleatoria en un 60% para el entrenamiento y un 40% para la prueba. La red se entrena con los mismos pesos iníciales, a través de cada uno de los 15 grupos. Los resultados con un área bajo la curva ROC mayor a 0.85 son mostrados en las figuras 40 y 41. Cada grupo de dos barras es una iteración (de entrenamiento y prueba), donde se comparan la sensibilidad, la especificidad y el AUC que es mostrada en la parte inferior (*el valor del AUC se muestra en negro sobre cada grupo de dos barras*). La sensibilidad mide la capacidad del sistema para detectar los casos en que se presenta la enfermedad (verdaderos positivos), mientras la especificidad mide la aptitud para detectar los casos en que no se sufre la enfermedad (verdaderos negativos). El AUC se utilizó antes, para describir la importancia de cada variable referente al diagnóstico real de la enfermedad, ahora se usa como medida de validación del sistema de clasificación. También, es importante tener en cuenta el número de pruebas, en las cuales se sobrepasa un área de 0.85, esto es, la cantidad de pares de barras en cada gráfica. Esta medida también cuantifica la validez de la clasificación.

En los resultados de la figura 40 se observa alrededor de 4 casos por conjunto, donde el AUC supera el umbral estipulado. En la mayoría de estos, la especificidad y la sensibilidad no presentan mayores diferencias, exceptuando la dirección vertical, que tiene unos valores apreciablemente distintos. Por otra parte, en la figura 41 se puede ver en promedio 5 casos que superan la validez del umbral 0.85. La sensibilidad y especificidad son más disparejos que los dados por la segmentación con transformada de Hough. De nuevo, en la dirección vertical se tiene el comportamiento más irregular, mientras en la dirección diagonal a 45°, se presentan 9 casos en los cuales el AUC es significativo.

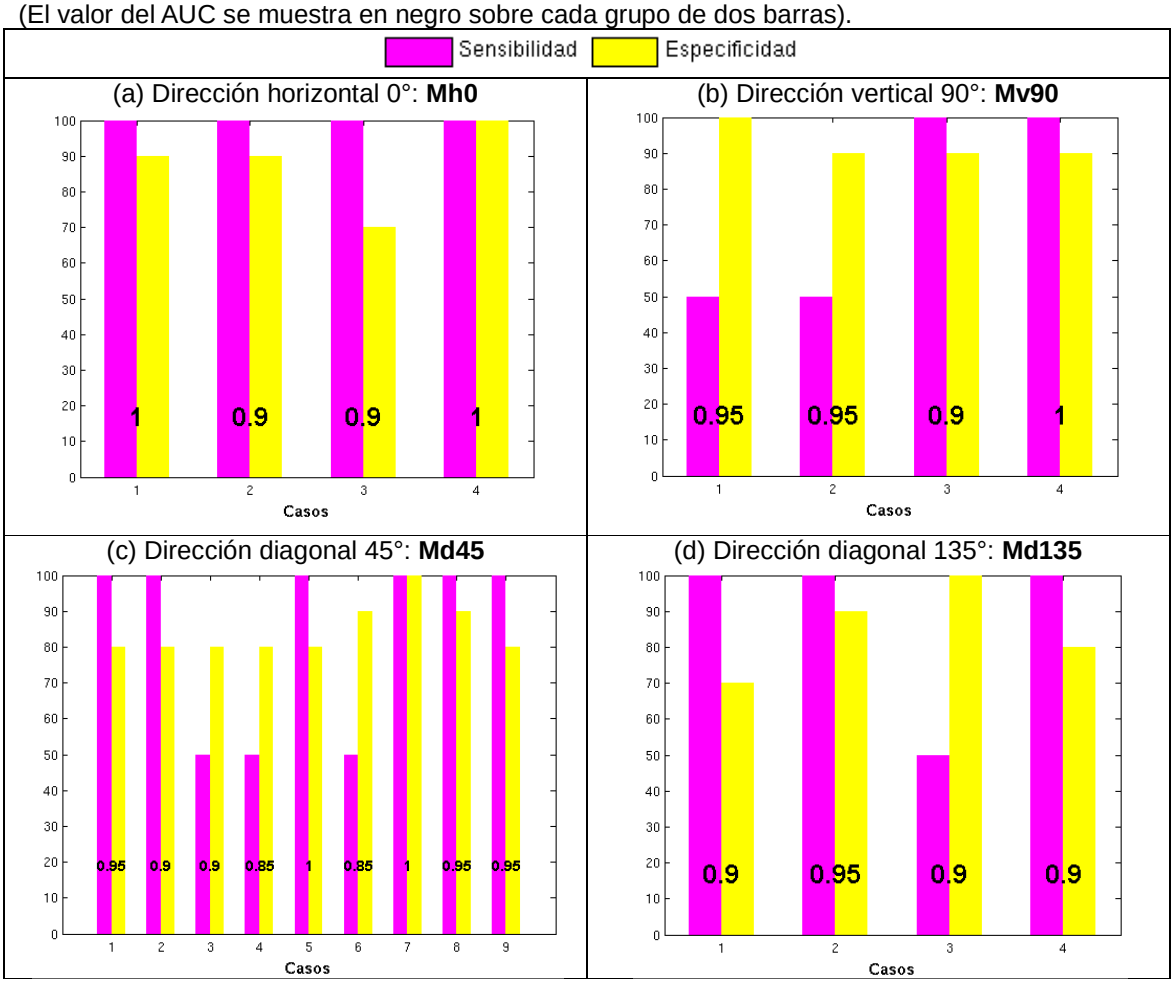
Figura 40. Sensibilidad, especificidad y AUC de la clasificación a través de la red neuronal RP para las variables que incluyen los descriptores termográficos de 2° orden, obtenidos a partir de diferentes direcciones de la matriz de co-ocurrencia, en las regiones segmentadas *mediante transformada de Hough* (El valor del AUC se muestra en negro sobre cada grupo de dos barras).



Fuente: el autor

La Tabla 16 resume el comportamiento de la RNA a través de todas las pruebas realizadas. Para los resultados con áreas mayores o iguales a 0.85, en las direcciones horizontales se tuvo una sensibilidad promedio del 100%, al igual que en la dirección diagonal a 45°, de la segmentación automática. Para estos valores seleccionados, el AUC oscila entre 0.9 y 0.99. Esto da un claro indicio de la excelente clasificación para estos casos. Las columnas que contienen todos los conjuntos (sin restricciones en el AUC), muestran el resultado general, dando un indicio del *sesgo* ocasionado por la pequeña muestra y el diminuto número de pacientes con cáncer. Por esto, en promedio, la especificidad es más alta que la

Figura 41. Sensibilidad, especificidad y AUC de la clasificación a través de la red neuronal RP para las variables que incluyen los descriptores termográficos de 2° orden, obtenidos a partir de diferentes direcciones de la matriz de co-ocurrencia, en las regiones segmentadas manualmente. (El valor del AUC se muestra en negro sobre cada grupo de dos barras).

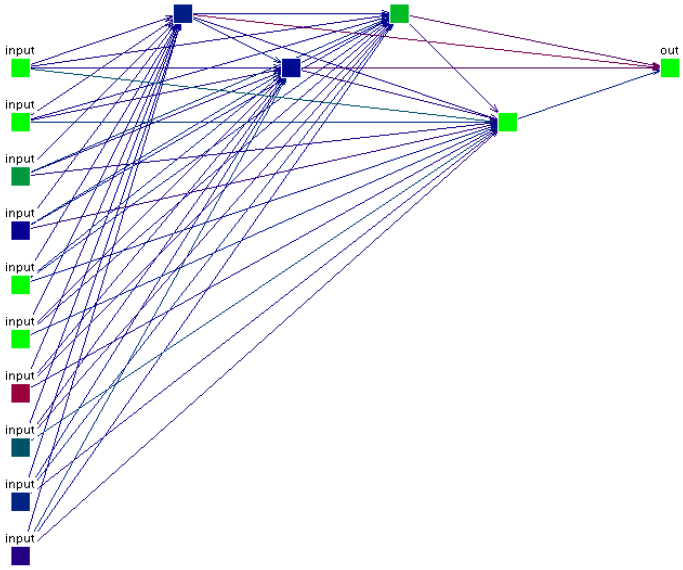


Fuente: el autor

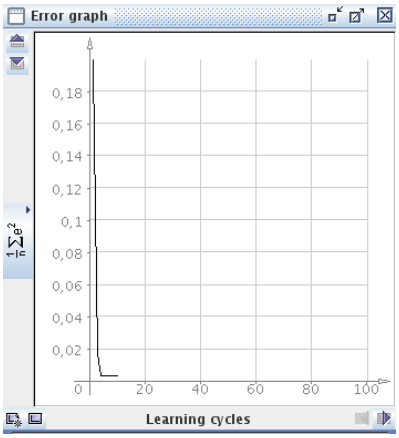
Tabla 16. Sensibilidad, especificidad y AUC promedios, en los datos con área significativa y en todo el conjunto, para la validación de la red en la Figura 38

Conjunto	Casos donde AUC > 0.85			Todos los casos		
	Sensibilidad	Especificidad	AUC	Sensibilidad	Especificidad	AUC
Hh0	100,00	85,00	0,99	56,67	80,67	0,72
Hv90	83,33	81,67	0,93	56,67	88,67	0,79
Hd45	100,00	82,50	0,95	60,00	79,33	0,71
Hd135	87,50	75,00	0,90	53,33	80,00	0,66
Mh0	100,00	87,50	0,95	56,67	82,00	0,69
Mv90	75,00	92,50	0,95	56,67	89,33	0,75
Md45	83,33	84,44	0,93	66,67	86,00	0,82
Md135	87,50	85,00	0,91	50,00	88,00	0,66

Figura 42. Ejemplo de una red neuronal entrenada mediante el algoritmo de Correlación en Cascada. En este caso se utilizo un conjunto de prueba Hh0.



(a) Arquitectura de la red neuronal con Correlación en Cascada, con 10 entradas, 4 neuronas ocultas añadidas y 1 neurona de salida.



(b) Gráfica del error cuadrático medio durante el entrenamiento de la red CC.

Fuente: el autor

sensibilidad durante todos los ensayos de este proceso. Es interesante que en general, para la dirección a 45°, el área bajo la curva ROC sea significativamente alta, con la segmentación manual.

7.2.2 Correlación en cascada

La redes neuronales feedfordward entrenadas con el algoritmo de Backpropagation (o sus modificaciones), dependen del número de neuronas ocultas. Estas se determinan, en la mayoría de los casos, mediante ensayos de prueba y error. En muchas ocasiones, este trabajo se vuelve tedioso y engorroso. Por esto, en esta investigación se decidió realizar el análisis con la arquitectura de redes neuronales en Correlación en Cascada. Esta metodología, discutida en la sección 1.4.1.6, automáticamente añade neuronas a la capa oculta, a través de la maximización de la correlación entre la salida de la neurona agregada y el error de toda la red.

Tabla 17. Parámetros del algoritmo de correlación en cascada, utilizado.

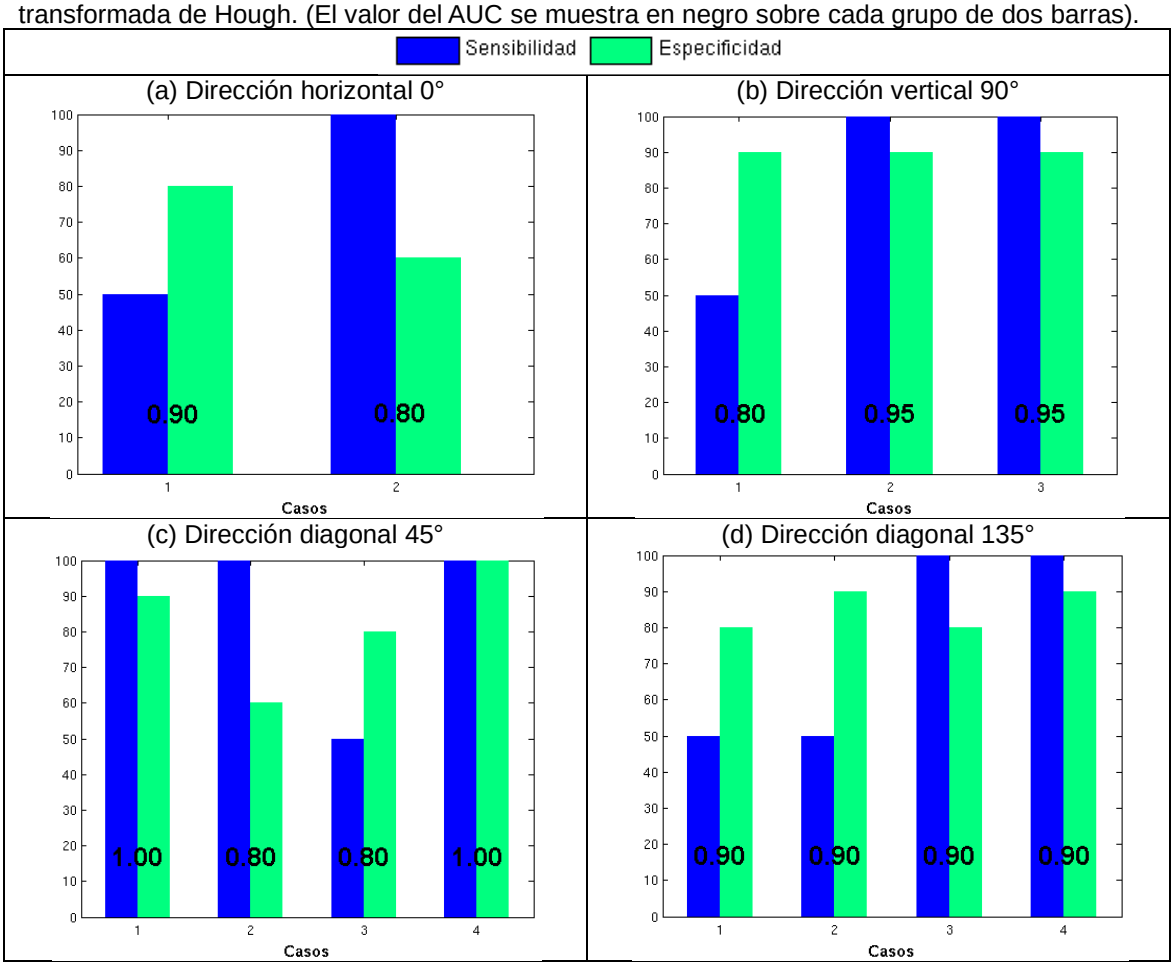
Fase	Parámetro	Valor	Descripción
General	Máximo error de salida de una neurona	0,2	El entrenamiento se aborta si el error es inferior
Salida	Cambio en el error	0,01	Fracción de cambio respecto al error anterior
	Paciencia de salida	50	Pasos a esperar con el cambio en el error
	Máximo de iteraciones	200	Número máximo de iteraciones, antes de empezar una nueva cascada
Candidatas	Cambio mínimo en la covarianza	0,04	Fracción de cambio respecto a la covarianza anterior
	Paciencia del candidato	25	Pasos a esperar con el cambio en la covarianza
	Máximo de actualizaciones de la covarianza	200	Máximo de actualizaciones antes de empezar una nueva cascada
	Máximo de neuronas candidatas	8	Tamaño de la capa de las candidatas
	Función de activación	tanh	Función de activación para los candidatos
Entrenamiento	Algoritmo de aprendizaje	BP	Función para entrenar a las candidatas
	η1	0.1	Paso descendente
	μ1	0.1	Momentum descendente
	η2	0.1	Paso ascendente
	μ2	0.1	Momentum ascendente
	c	0.1	Factor de eliminación de pedazos planos
	Cascadas	10	Máximo número de cascadas

7.2.2.1 Pruebas para la red neuronal con correlación en cascada

La red que se usó posee 10 entradas y una salida. La neurona de salida tiene una función de trasferencia (o activación) tangente hiperbólica. En la Figura 42 se puede ver un ejemplo de la arquitectura CC utilizada y su respectiva gráfica de error. Los parámetros del algoritmo se muestran en la Tabla 17. El software utilizado para realizar las simulaciones con el algoritmo de Correlación en Cascada, fue Java Neural Network Simulator (JNNS[24]). Este es un versátil programa de código libre, que implementa numerosas técnicas de RNA. Nuevamente se utilizó la división de los datos formulada antes para la red BP: 60% para entrenar y 40% para probar. Se planteó utilizar los ocho conjuntos de variables para generar las sensibilidades, especificidades y áreas bajo la curva

[24] Disponible en www.ra.cs.uni-tuebingen.de/SNNS

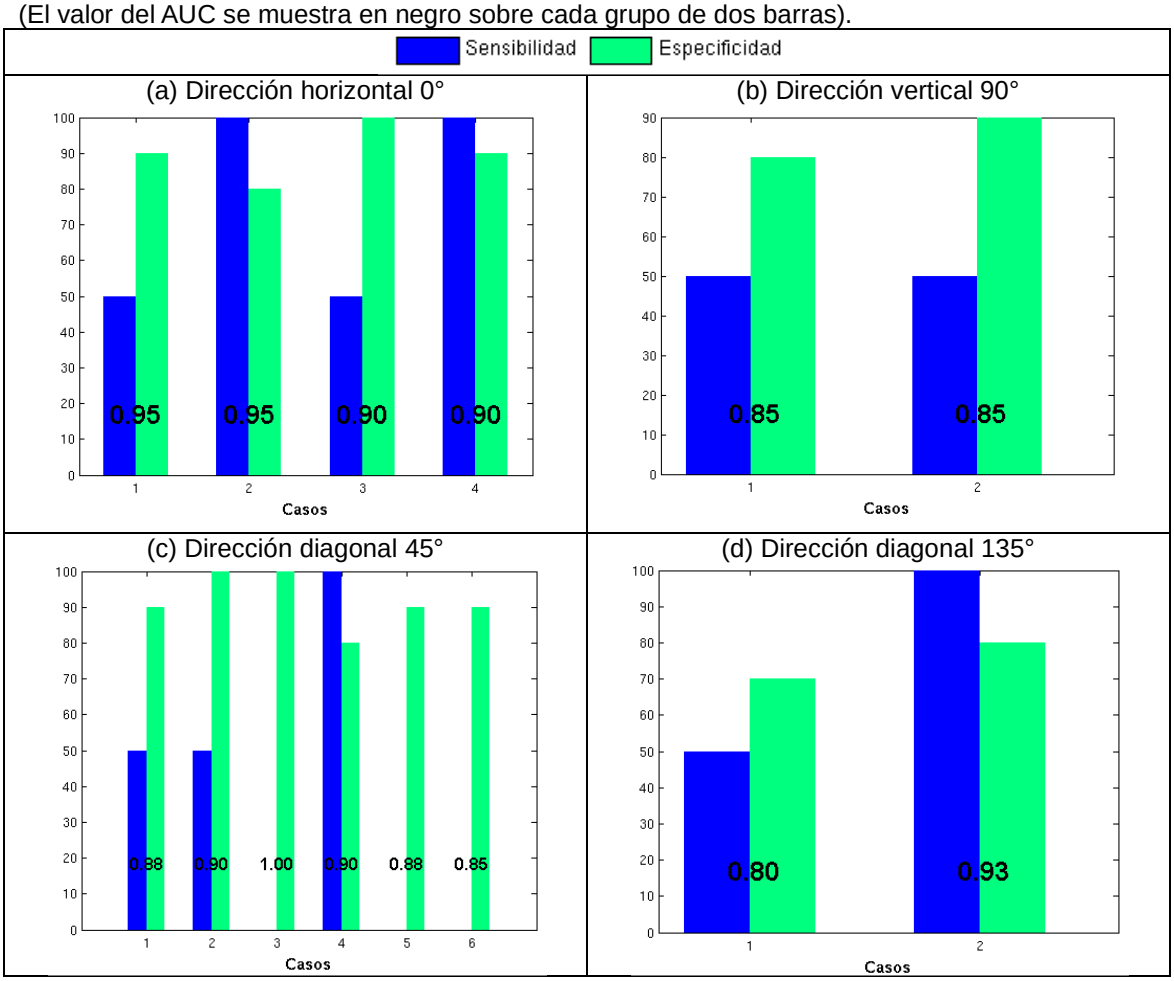
Figura 43. Sensibilidad, especificidad y AUC de la clasificación a través de la red neuronal CC para las variables que incluyen los descriptores termográficos de 2° orden, obtenidos a partir de diferentes direcciones de la matriz de co-ocurrencia, en las regiones segmentadas mediante transformada de Hough. (El valor del AUC se muestra en negro sobre cada grupo de dos barras).



Fuente: el autor

ROC, con el fin de analizar el comportamiento del algoritmo CC respecto a la presencia o ausencia del carcinoma con glándula mamaria. En las Figuras 43 y 44 se ilustran los parámetros mencionados anteriormente, para los conjuntos de variables de la segmentación manual y de la segmentación con transformada de Hough. Ambas ilustraciones muestran un bajo número de casos con un AUC mayor a 0.8, incluso dentro de estas selecciones, en algunas ocasiones la sensibilidad llega a ser muy baja.

Figura 44. Sensibilidad, especificidad y AUC de la clasificación a través de la red neuronal CC para las variables que incluyen los descriptores termográficos de 2° orden, obtenidos a partir de diferentes direcciones de la matriz de co-ocurrencia, en las regiones segmentadas manualmente. (El valor del AUC se muestra en negro sobre cada grupo de dos barras).



Fuente: el autor

Tabla 18. Sensibilidad, especificidad y AUC promedios, en los datos con área significativa y en todo el conjunto, para la validación de la red CC

Conjunto	Casos donde AUC > 0.85			Todos los casos		
	Sensibilidad	Especificidad	AUC	Sensibilidad	Especificidad	AUC
Hh0	75,00	70,00	0,85	43,33	76,67	0,56
Hv90	83,33	90,00	0,90	50,00	86,67	0,64
Hd45	87,50	82,50	0,90	50,00	87,33	0,69
Hd135	75,00	85,00	0,90	46,67	84,67	0,63
Mh0	75,00	90,00	0,93	56,67	87,33	0,69
Mv90	50,00	85,00	0,85	40,00	89,33	0,58
Md45	33,33	91,67	0,90	33,33	92,00	0,76
Md135	75,00	75,00	0,86	36,67	85,33	0,64

Al respecto, la Tabla 17 resume el comportamiento general, para áreas significativas (mayores a 0.8) y en todos los casos. En esta, para la mayoría de las direcciones se ve un AUC promedio bastante baja, cercana a 0.5. El mayor valor de área, ocurre para los datos en la segmentación manual con la matriz de co-ocurrencia en dirección noroeste, pero con una sensibilidad del 33%.

7.2.3 Red neuronal probabilística (PNN[25])

Las redes probabilísticas son un tipo de redes neurales de base radial, compuesta de cuatro capas. Estas entrenan rápidamente, pues el aprendizaje es hecho en un solo paso para cada vector de entrenamiento. Las PNN estiman la función de densidad de probabilidad para cada clase según la muestra de entrenamiento, a través de la comparación de la distancia euclidiana del vector de prueba con cada vector de entrenamiento, normalizando las distancias a través de una función de base radial (por cada categoría) y sumando las contribuciones para cada clase [73].

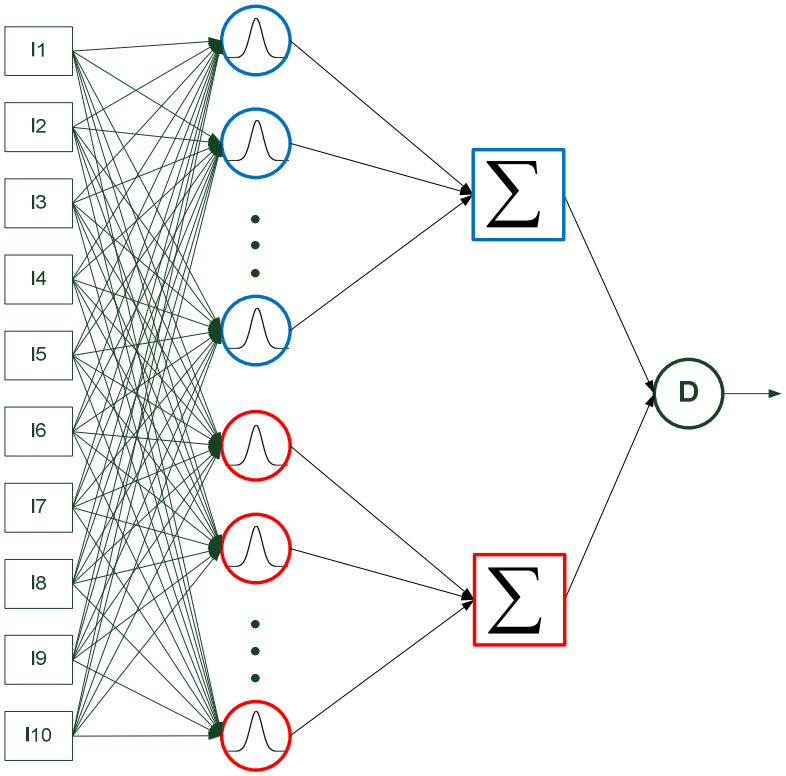
A continuación se describe la actividad de cada capa dentro de la PNN:

1. Capa de entrada: esta compuesta de las variables que alimentan a la capa de patrones.

2. Capa de patrones: esta capa tiene una neurona por cada caso en el conjunto de entrenamiento. Cuando se presenta un vector proveniente de la capa de entrada, una neurona calcula la distancia euclidiana del caso de prueba respecto al punto central de la misma (caso de entrenamiento) y luego se le aplica una función de base radial con un ancho determinado. El resultado pasa a la capa categórica.

[25] En inglés, Probabilistic Neural Network

Figura 45. Red Neuronal Probabilística utilizada para la clasificación del carcinoma de mama.

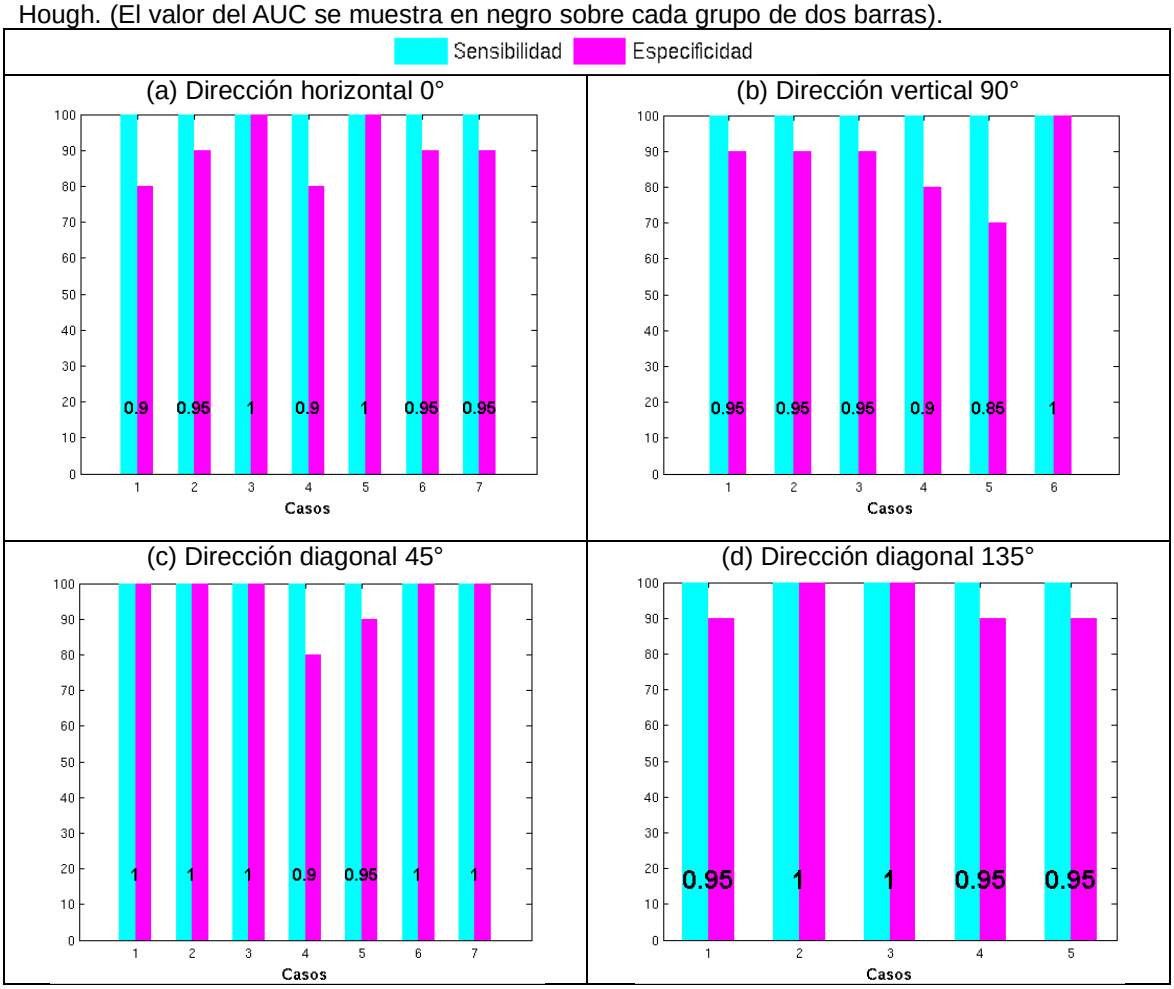


Fuente: el autor

3. Capa categórica o de suma: contiene una neurona por cada clase de la variable categórica de salida. Las neuronas de la capa de patrones se conectan a las neuronas categóricas correspondientes, es decir, se relaciona cada conjunto de entrenamiento a su respectiva clase. Posteriormente, se suman las entradas, produciendo una salida por cada nivel de clasificación.

4. Capa de decisión: compara las salidas de la capa categórica y selecciona el máximo valor para predecir la clase a la cual pertenece el vector de entrada.

Figura 46. Sensibilidad, especificidad y AUC de la clasificación a través de la red neuronal PNN. Los descriptores termográficos provienen de las regiones segmentadas mediante transformada de Hough. (El valor del AUC se muestra en negro sobre cada grupo de dos barras).

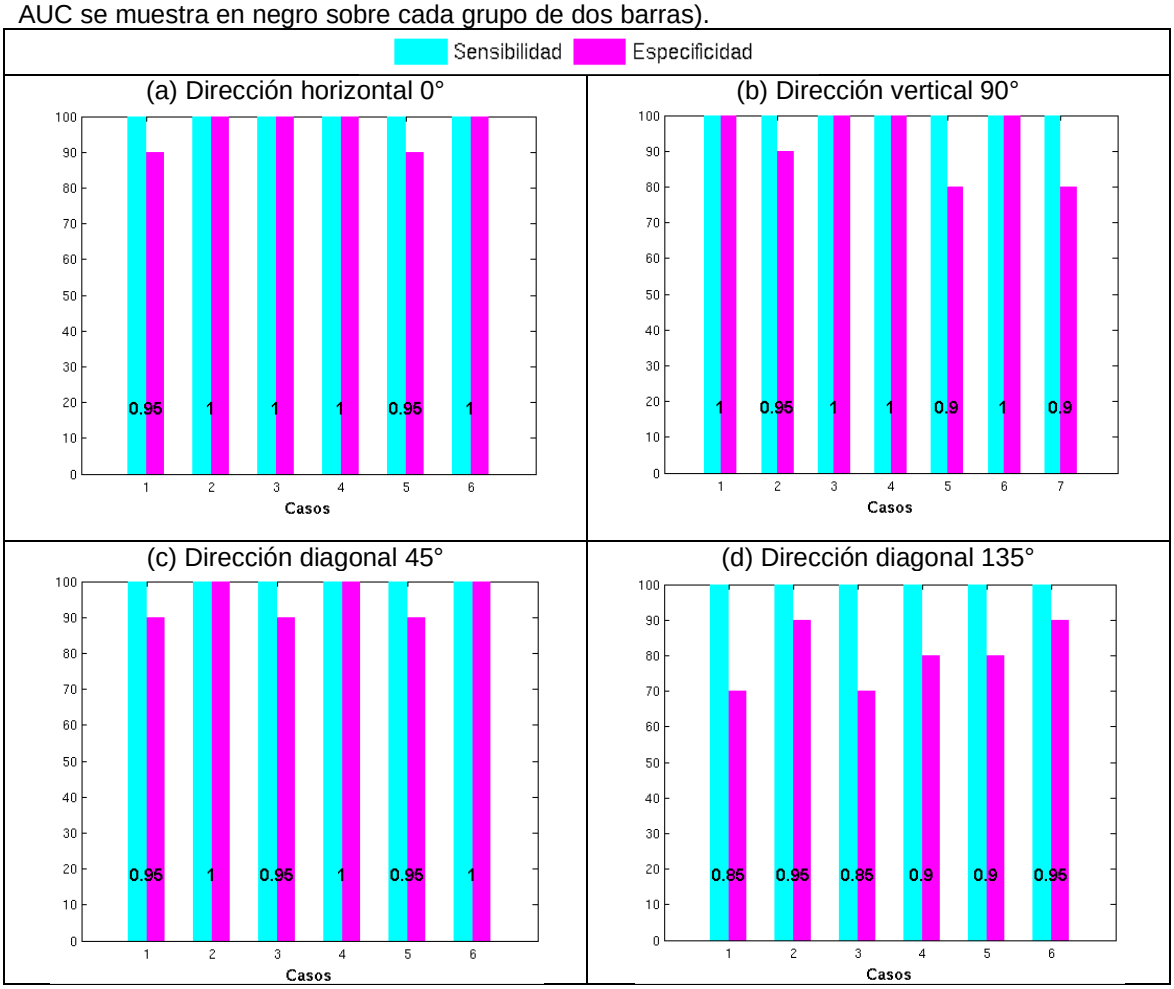


Fuente: el autor

Tabla 19. Sensibilidad, especificidad y AUC promedios, en los datos con área significativa y en todo el conjunto, para la validación de la PNN

Conjunto	Casos donde AUC > 0.85			Todos los casos		
	Sensibilidad	Especificidad	AUC	Sensibilidad	Especificidad	AUC
Hh0	100,00	90,00	0,95	76,67	87,33	0,82
Hv90	100,00	86,67	0,93	70,00	90,67	0,80
Hd45	100,00	95,71	0,98	63,33	97,33	0,80
Hd135	100,00	94,00	0,97	66,67	94,67	0,81
Mh0	100,00	96,67	0,98	66,67	97,33	0,82
Mv90	100,00	92,86	0,96	70,00	96,67	0,83
Md45	100,00	95,00	0,98	66,67	97,33	0,82
Md135	100,00	80,00	0,90	73,33	86,67	0,80

Figura 47. Sensibilidad, especificidad y AUC de la clasificación a través de la red neuronal PNN. Los descriptores termográficos provienen de las regiones segmentadas manualmente (el valor del AUC se muestra en negro sobre cada grupo de dos barras).



Fuente: el autor

La Figura 45 muestra la PNN utilizada en este trabajo para clasificar la información de las pacientes, en cáncer y no cáncer. Como se nota, en la capa de patrones se distingue las neuronas de los casos con y sin carcinoma, correspondientes a los colores rojo y azul. Las funciones gaussianas utilizadas, poseen una desviación de 0.18, determinada con pruebas sucesivas.

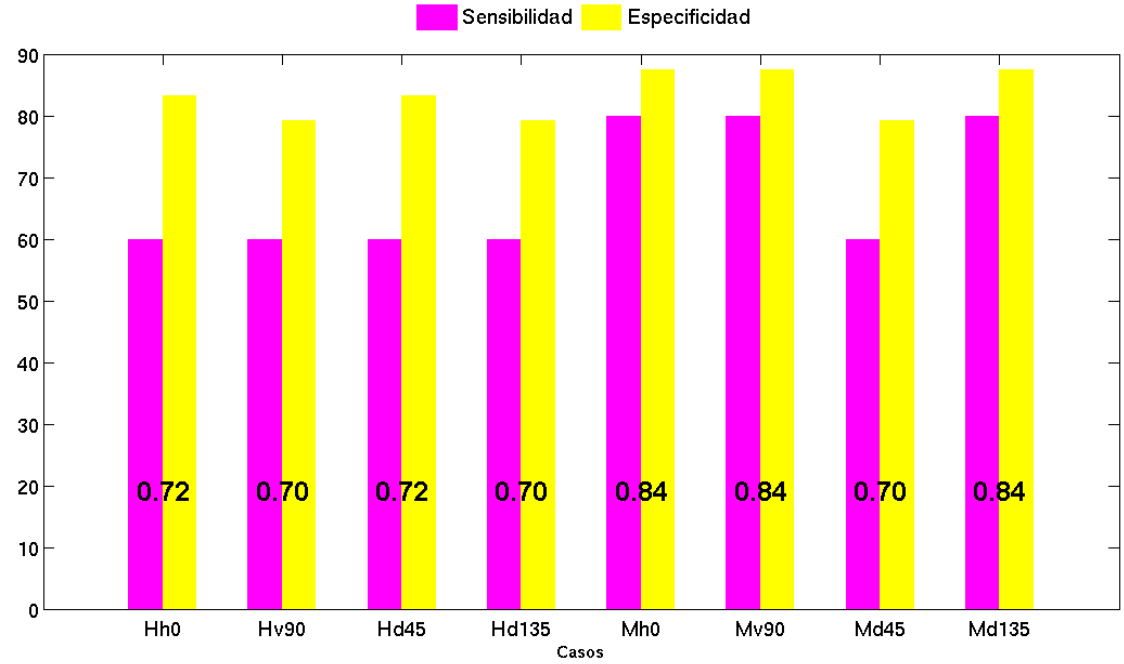
7.2.3.1 Pruebas para la red neuronal probabilística

Al igual que las pruebas para los casos anteriores, en la PNN se utilizaron divisiones aleatorias en porcentajes de 60 y 40% para el entrenamiento y la prueba, respectivamente. En la Figura 46 se representan la sensibilidad, la especificidad y la AUC, de forma similar a los diagramas de barras anteriores, de la segmentación por medio de la transformada de Hough. En la Figura **47** se emplean los mismos parámetros para la segmentación manual. Los resultados de los valores verdaderos positivos, verdaderos negativos y área bajo la curva ROC, son mucho mejores que los mostrados en Figura 41, para ambos tipos de segmentación. En promedio, se presentaron 6 casos donde el AUC supera a 0.85. Esto resalta el excelente comportamiento del algoritmo en la clasificación. Es evidente la similaridad en todos los resultados respecto a la sensibilidad y especificidad, donde en varias ocasiones se alcanzan picos de 100% para ambos parámetros. La Tabla 19 resume las pruebas realizas con la PNN. Se observa una sensibilidad del 100% en todos los conjuntos seleccionados con la AUC. Por otro lado, el resultado general (AUC < 1), muestra un área muy significativa (mayor a 0.85), para ser una cantidad promedio de 15 pruebas en cada una de las direcciones, con una ligera ventaja para los casos provenientes de la segmentación manual.

7.2.4 Fuzzy c - means

Los algoritmos utilizados anteriormente para clasificar, son supervisados, por cuanto dependen del entrenamiento o aprendizaje. Además, la muestra poblacional de pacientes es muy limitada, lo que sugiere el empleo de técnicas de agrupamiento (en general cualquiera no supervisada) para encontrar los dos clúster (paciente enfermo y paciente sano) y posteriormente compararlos respecto a la verdadera asociación, proveniente de los resultados de la biopsia. De esta forma, se aprovecha todo el conjunto de datos para encontrar una posible clasificación.

Figura 48. Sensibilidad, especificidad y AUC de la clasificación con Fuzzy C – means. (el valor del AUC se muestra en negro sobre cada grupo de dos barras).



Fuente: el autor

Fuzzy C – means, es una técnica de agrupamiento, descrita en la sección 1.4.2, que encuentra las membresías de los datos respecto a los grupos, maximizando la distancia entre los centros de los clúster y minimizando la distancia entre los puntos y el centro del grupo más cercano. FCM se aplica en este trabajo, con los siguientes valores en los parámetros:

- Exponente de la matriz de membresía, $m = 2$.
- Número máximo de iteraciones igual a 100.
- Umbral mínimo de cambio en la función objetivo, igual a 10^{-5}.

Dado que no existe una partición en los datos para entrenamiento y prueba, el algoritmo FCM se aplicó una sola vez, por cada conjunto de datos (direcciones: horizontal, vertical y diagonales). El algoritmo de Fuzzy C – means creó dos clúster: uno que aglomera los casos con carcinoma y otro que agrupa los casos de pacientes sanas. La asociación de los grupos correspondientes, se determinó por

la mayor cantidad de aciertos en la clasificación (incluyendo positivos y negativos) y por el máximo valor de la membresía.

La Figura 48 es un diagrama de barras, que muestra los resultados para cada uno de los grupos con las variables seleccionadas. La especificidad para la mayoría de los casos es del 80%, la sensibilidad es alrededor del 60% y el área bajo la curva ROC tiene un máximo de 0.84. Este último valor indica que el rendimiento del sistema de clasificación es aceptable, puesto que el AUC es una forma de medir el desempeño de un sistema de clasificación (un área de 0.5, se atribuye a un proceso de clasificación aleatorio). En especial, la segmentación manual en tres de las cuatro direcciones tiene el AUC en 0.84. Se destaca que el algoritmo se hizo empleando toda la muestra y no porciones de ella, como ocurre en el aprendizaje supervisado.

7.2.5 Comparación de los algoritmos de clasificación

Las técnicas utilizadas para la clasificación dicotómica del carcinoma de glándula mamaria han involucrado en su mayoría algoritmos supervisados de redes neuronales artificiales y una metodología no supervisada: el algoritmo de agrupamiento *fuzzy c-means.* En cada una de las pruebas se empleó el mismo tipo de división de la muestra: aleatorio y repetitivo, para solventar el número reducido de pacientes que participaron en el estudio.

La Tabla 20 compara todos los algoritmos de clasificación empleados, con los promedios generales, a través de las mejores pruebas y todos los conjuntos. La comparación se realiza a través de la selección de las pruebas con área bajo la curva ROC por debajo de un umbral. Sin embargo, la técnica de FCM siempre actúa con todos los datos. Para las áreas significativas, el peor comportamiento lo presenta la Correlación en Cascada, con una AUC del 89%, y una sensibilidad alrededor del 69%.

Tabla 20. Sensibilidad, especificidad y AUC promediadas a través de todas las pruebas y direcciones para cada algoritmo de clasificación.

Algoritmo	Sensibilidad	Especificidad	AUC
BP	89,58	84,20	0,94
CC	69,27	83,65	0,89
PNN	100,00	91,36	0,96
FCM	67,50	83,33	0,75

Es interesante, que en el análisis con FCM se determina unas especificidad, sensibilidad y área, relativamente altas, especialmente, si se tiene en cuenta, que es un análisis con todos los datos disponibles.

8 CONCLUSIONES Y RECOMENDACIONES

La investigación realizada permite evaluar la incidencia de factores de riesgo de tipo clínico en combinación con variables adquiridas a partir de los termogramas frontales de mama, en la presencia del carcinoma de glándula mamaria. La metodología propuesta por este trabajo, se puede extender a otros estudios que involucren un análisis similar de variables, especialmente los que impliquen a la termografía infrarroja como técnica diagnóstica.

A continuación se relacionan algunas conclusiones importantes del trabajo realizado y algunas recomendaciones para futuros estudios:

Se desarrolló un protocolo para el registro termográfico de mama, que facilita la adquisición de la información térmica, al evitar factores que pueden ocasionar la distorsión de los datos, debido a un manejo inadecuado, tanto del recinto como de la preparación de paciente. De esta forma, las imágenes térmicas de las pacientes reales que se tomaron a lo largo de la investigación, fueron obtenidas con una calidad que permite el análisis de las variables, sin repercusiones para el desarrollo de los algoritmos.

A través de los formularios se recopiló la información sociodemográfica, hereditaria, hormonal y fisiológica. Estos bio-datos fueron almacenados en una base de datos con el fin de ordenar el proceso y posteriormente, permitir un análisis coherente. A continuación, se preseleccionaron los factores de riesgo mediante dos criterios: (1) la cantidad de información disponible respecto a cada variable, de la población que participó en el estudio, (2) la incidencia de cada factor respecto al cáncer, en diferentes investigaciones epidemiológicas importantes. Este es el primer paso para la selección de las variables que sugieren la presencia o ausencia del cáncer de mama.

Se propuso una metodología para la segmentación de los senos en la termografía frontal de mama. Al evaluar el proceso, se logró una efectividad del 74% con una desviación estándar del filtro gaussiano, empleado en la detección de bordes, de 1.8. Sin embargo, se alcanza una efectividad hasta del 92%, al variar este parámetro manualmente. De esta forma, los resultados experimentales corroboran la eficacia de sustentar la segmentación de las glándulas mamarias en imágenes termográficas basándose en las características geométricas de su contorno. Es decir, en la forma parabólica que estos evidencian una vez se somete el termograma a un proceso de detección de bordes utilizando el operador Canny, y en identificar estas formas características por medio de la Transformada Parabólica de Hough.

De forma análoga a la segmentación automática, se realizó una segmentación manual, con forma elíptica, con el objeto de generar una región más específica de los senos. Esto permite una comparación entre dos modelos geométricos diferentes: el parabólico con fusión de contornos, que contiene una porción mayor de las mamás, llegando incluso hasta las axilas; el elíptico, que es más selectivo y congrega la mayor parte visual y anatómica de las glándulas mamarias.

La selección de las variables fusionadas, se realizó utilizando tres métodos distintos: coeficientes de correlación, área bajo la curva ROC y selección secuencial de características. Los tres métodos posicionan a la *Edad* como la variable con mayor significancia respecto al resultado histopatológico. Seguido, se encuentra *la edad de la menopausia*, y las variables termográficas *media, homogeneidad, curtosis, energía* (1° y 2° orden), *moda y máximo*. Sin embargo, debido al tamaño de la muestra y a la presencia de variables categóricas, se optó por escoger el AUC con 10 variables, como el criterio de selección para la posterior clasificación.

A partir de la selección de las variables fusionadas, divididas en 8 conjuntos, uno

por cada dirección de la matriz de co-ocurrencia, se implementaron cuatro algoritmos de clasificación para analizar la relación existente entre el carcinoma de mama y las respectivas características: las redes neuronales artificiales Feedfordward Backpropagation, correlación en cascada y probabilística, y la técnica de agrupamiento Fuzzy C – Means. Los tres primeros, se entrenaron y validaron 15 veces con divisiones aleatorias de la muestra, mientras FCM utilizó una vez la población total de cada grupo de datos. Esta metodología de análisis se aplicó debido a que las pacientes que participaron en el estudio fueron 29, con resultado de biopsia.

Se determinó la sensibilidad, especificidad y área bajo la curva ROC, de cada algoritmo de clasificación, por cada prueba y grupo de datos. De acuerdo con estos parámetros, la red neuronal probabilística logra los mejores resultados, alcanzando en varias ocasiones una sensibilidad del 100% y una especificidad de hasta el 96%, para las pruebas seleccionadas. Desde un punto de vista más general, el AUC promedio de esta arquitectura, supera el valor de 0.8. Una cantidad excepcional para la pequeña población de estudio. Por otro lado, el algoritmo de FCM, presenta un área promedio de 0.75, llegando en algunas situaciones a 0.84. Las sensibilidades medias de todas las pruebas siempre son menores que las especificidades, debido a que el 83% de la muestra, son casos en los que no se presenta cáncer.

A través de todas las pruebas, tanto para la segmentación manual como para la segmentación mediante transformada parabólica de Hough, el análisis ROC no muestra una marcada diferencia en las implementaciones de la matriz de co-ocurrencia a través de las cuatro direcciones, por lo tanto, ninguna dirección es dominante para generar los descriptores termográficos.

Aunque los resultados logrados son optimistas, para la fusión de variables clínicas y termográficas, en la detección del carcinoma de glándula mamaria, la muestra

de pacientes es muy pequeña para decidir con certeza el uso del diagnóstico de mama con la técnica propuesta. Por tanto, es trascendental lograr aumentar la cantidad de pacientes que participan en el estudio, con el fin de alcanzar un grado de validación mayor.

Para una futura continuación de este trabajo, realizar un algoritmo adaptativo de la desviación estándar del filtro gaussiano, podría permitir alcanzar una efectividad cercana al 100% en la segmentación de las mamas, con base en una forma parabólica. Al respecto, la forma geométrica elíptica, es una alternativa válida para una selección más minuciosa de los senos.

El principal inconveniente de esta investigación, fue la adquisición de la información, por cuanto este proceso involucra una participación activa de los médicos, en especial de toda la gestión que pueden realizar para la consecución de la muestra poblacional. Por esto, se recomienda que al iniciar este tipo de trabajos, que involucran una técnica novedosa que no tiene antecedentes en la región o incluso en el país, se fije desde el inicio y se verifique, un aporte continuo y permanente de toda la gestión humana y médica.

La metodología propuesta a lo largo de este trabajo, se puede implementar de nuevo cuando la muestra aumente significativamente. De esta forma, la selección de variables secuencial, empleando regresión logística, se utilizaría con un mayor rendimiento al disponer de más datos. Por ejemplo, realizando entrenamientos y pruebas, semejantes a los algoritmos de entrenamiento supervisados. El área bajo la curva ROC, sigue siendo un excelente criterio de selección, pero también uno de los mejores métodos para evaluar el desempeño del sistema de clasificación.

De acuerdo a los resultados de la red probabilística y el algoritmo FCM, en varias pruebas la segmentación manual alcanza una validez alta, con sensibilidades y

especificidades coherentes. Por tanto, se recomienda elaborar un algoritmo de segmentación automática que detecte formas geométricas elípticas, como la transformada de Hough elíptica.

El alcance de este proyecto, se limitó a trabajar con la termografía de mama frontal, aunque los resultados obtenidos son más que aceptables, involucrar las termografías laterales y oblicuas, podría aumentar la capacidad de clasificación y lograr índices de desempeño muy superiores a los alcanzados.

9 BIBLIOGRAFÍA

[1] Ferlay J, Bray F, Pisani P, Parkin DM. GLOBOCAN 2002: cancer incidence, mortality and prevalence Worldwide IARC, Cancer base No. 5. versión 2.0, IARC press, Lyon, 2004.

[2] Gerald C. Holst. Common Sense Approach to thermal imaging. Publication of SPIE. Florida, Usa. JCD Publishing. 2000.

[3] Jones, B.F. A reappraisal of the use of infrared thermal image analysis in medicine: Medical Imaging, IEEE Transactions on Volume 17, Issue 6, Dec. 1998 Page(s):1019 – 1027.

[4] Lawson RN, Chughtai MS. Breast cancer and body temperatures. Can Med Assoc J. 88:68-70, 1963.

[5] Gautherine M, gros CM. Breast thermography and cancer risk prediction. Cancer 45:51-56, 1980.

[6] Stark AM. The value of risk factors in screening for breast cancer. Eur J Cancer 11:147-150, 1985.

[7] Feig SA, Shaber GS, Schwartz Gf et al. Thermography, mamography, and clinical examination in breast cancer screening. Radiology 122:123-127, 1977.

[8] Keyserling JR, Ahlgen PD, Yu E, Belliveau, M.YassA. Functional Infrared Imaging of the Breast. IEEE Eng Med Biol Mag. 2000,19(3): 30-41

[9] Elliot RL, Head JF, and Werneke DK. Thermography in breast cancer: Comparison with patient survival, TNM classification and tissue ferritin concentration. Proc Amer Soc Clin Oncol, vol. 9, pp. 277, 1991.

[10] Head JF, Wang F, Lipari CA, Elliot RL. The important role of infrared imaging in breast cancer. IEEE Eng Med Biol Mag, vol. 19, pp. 52-57, 2000.

[11] Head JF, Lipari CA, Wang F, Davidson JE, Elliot RL. Application of second generation of infrared imaging and computerized image analysis to breast cancer risk assessment. Proc 18th Ann Int Conf IEEE Eng Med Biol Soc, pp.

1019-1021, 1996.

[12] Qi H, Head JF. Assymetry analysis using automatic segmentation and classification for breast cancer detection in thermograms. Proc 23th Ann Int Conf IEEE Eng Med Biol Soc, 2001.

[13] J. Koay, C. Herry, M. Frize. Analisys of Breast Thermography with Artificial Neural Network. Proc 26th Ann Int Conf IEEE Eng Med Biol Soc, 2004

[14] Parisky, Y.R. Efficacy of computerized infrared imaging análisis to evaluate mammographically susupicious lesions. American Roentgen Society. January 2003

[15] J.R. Keyserlingk, P.D. Ahlgren, E. Yu, N. Belliveau, and M. Yassa, “Functional infrared imaging of the breast," *IEEE Eng. Med. Biol. Mag.,* vol. 19, pp. 30-41, 2000.

[16] J.F. Head, F. Wang, C.A. Lipari, and R.L. Elliott, “The important role of infrared imaging in breast cancer: New technology improves applications in risk assessment, detection, diagnosis and prognosis,” *IEEE Eng. Med. Biol. Mag.,* vol. 19, pp. 52-57, 2000.

[17] Chummy S. Sinnatamby, R. J. Last, Anatomía de Last, Paidotribo Editorial, 2006.

[18] Junceda Avello Enrique, Oviedo , Cáncer de Mama,: Universidad de Oviedo. Servicio de Publicaciones, D.L. 1988

[19] Ramsay et. al., Anatomy of the lactating human breast redefined with ultrasound imaging. s.l. : J. Anat. 206:525-34.

[20] Anderson, MD. *¿QUÉ ES EL CÁNCER?.* s.l. : La Fundación M.D. Anderson España - Centro Oncológico. Consultado, 19 de Diciembre.2007.

[21] Anatomía y Fisiología del Aparato Reproductor Masculino y Femenino (en español). s.l. : Editorial CEP, 19 de diciembre, 2007.

[22] Netter H. Frank, Interactive Atlas of Clinical Anatomy,1997.

[23] Merlin JL, Barberi-Heyob M & Bachmann N. *In vitro* comparative evaluation of trastuzumab (Herceptin®) combined with paclitaxel (Taxol®) or docetaxel

(Taxotere®) in Her-2-expressing human breast cancer cell lines. Annals of Oncology 2002; 13: 1743-1748.

[24] Peto J, Mack TM. High constant incidence in twins and other relatives of women with breast cancer. Nat Genet 2000; 26: 411-414.

[25] Pike MC, Spicer DV, Dahmoush L, et al. Estrogens, progestagens, normal breast cell proliferation and breast cancer risk. Epidemiol Rev 1993; 15:17-35.

[26] Kelsey JL, Fischer DB, Holford TR, et al. Exogenous estrogens and other factors in the epidemiology of breast cancer. J Natl Cancer Inst 1981; 67:327-333.

[27] National Center for Health Statistics. Seer cancer statistics review, 1973-1995. Bethesda, MD: U.S. National Cancer Institute, 1998.

[28] Sturgeon SR, Schairer C, Gail M, et al. Geographic variation in mortality from breast cancer among white women in the United States. J Natl Cancer Inst 1995; 87:1846-1853.

[29] Rosen PP, Groshen S, Kinne DW, et al. Factors influencing prognosis in node-negative breast carcinoma: analysis of 767 T1N0M0/T2N0M0 patients with long-term follow up. J Clin Oncol. 1993;11:2090–2100.

[30] Slamon DJ, Clark GM, Woung SG et al. Human breast cancer: correlation of relapse and survival with amplification of the Her-2/neu oncogen. Science 1987; 235: 177-182.

[31] Hsieh C-C, Trichopoulos D, Katsouyanni K, et al. Age at menarche, age at menopause, height and obesity as risk factors for breast cancer: associations and interactions in an international case-control study. Int J Cancer 1990; 46:796-800.

[32] Kelsey JL, Gammon MD, John EM. Reproductive factors and breast cancer. Epidemiol Rev 1993; 15:36-47.

[33] Brinton LA, Schairer C, Hoover RN, et al. Menstrual factors and risk of breast cancer. Cancer Invest 1988; 6: 245-254.

[34] Layde PM, Webster LA, Baughman AL, et al. The independent associations of parity, age at first full term pregnancy, and duration of breastfeeding with the risk of breast cancer. Cancer and Steroid Hormone Study Group. J Clin Epidemiol 1989; 42:963-973.

[35] Zheng T, Holford TR, Mayne ST, et al. Lactation and breast cancer risk: a case-control study in Connecticut. Br J Cancer 2001; 84:1472-1476.

[36] Collaborative Group on Hormonal Factors in Breast Cancer (Writing Committee: Beral V, Bull D, Peto R, et al). Breast cancer and breastfeeding: collaborative reanalysis of individual data from 47 epidemiological studies in 30 countries, including 50302 women with breast cancer and 96973 women without the disease. Lancet 2002; 360:187-195.

[37] Jernstrom H, Lubinski J, Lynch HT, et al. Breast-feeding and the risk of breast cancer in BRCA1 and BRCA2 mutation carriers. J Natl Cancer Inst 2004; 96:1094-1098.

[38] Jain, Anil K. Fundamentals of Image Processing. Englewood Cliffs: Prentice Hall, 1989. p.233.

[39] Jhon C. Russ, The Image Processing Handbook, 5 ed, CRC Press, 2007.

[40] R. Gonzales, R. Woods, Digital Image Processing, Third Edition, Pearson Prentice Hall, 2008.

[41] Linda G. Shapiro and George C. Stockman (2001): “Computer Vision”, pp 279-325, New Jersey, Prentice-Hall.

[42] N. Otsu, “A Threshold Selection Method from Gray – Level Histograms”, IEEE Trans. SMC-9:1, January 1979.

[43] Canny, J., *A Computational Approach To Edge Detection*, IEEE Trans. Pattern Analysis and Machine Intelligence, 8:679-714, 1986.

[44] Method and Means for Recognizing Complex Patterns. Hough and P.V.C. 1962. US pattent 3,069,654.

[45] Jafri, M.Z and Deravi, F. Efficient algorithm for detection of parabolic curves. Vision Goetry. III, 1994, 2356.

[46] R.M. Haralick, K. Shanmugam, I. Dinstein, Textural features for image classification, IEEE Trans. Syst. Man Cybern. 3 (6) (1973) 610–621.

[47] S. Theodoridis, K. Koutroumbas, Pattern Recognition, 3 edition, Academic Press, 2006.

[48] Trias R. inteligencia Artificial en medicina. Estado actual y perspectivas. Med Clin Barc 1993; 100 Supl 1:45-46.

[49] Bernal E. Inteligencia Artificial Aplicada al Diagnóstico Médico. Estado del arte Tesis de Maestría en Ingeniería, Ingeniería de Sistemas y Computación, Universidad Nacional de Colombia 2006. Disponible en: http://es.geocities.com/edwin99109/final_arte1.pdf

[50] Wu Y; Doi K; Giger M L; Nishikawa R M. Computerized detection of clustered microcalcifications in digital mammograms: applications of artificial neural networks. Medical physics 1992; 19(3):555-60.

[51] Lo J. Y., Baker J. A., Kornguth P. J. and Floyd C. E. Effect of patient history data on the prediction of breast cancer from mammographic findings with Artificial Neural Networks. Acad Radiol 1999; 6:10-15.

[52] E. Y. K Ng, and E. C. Kee, “Advanced integrated technique in breast cancer thermography, Journal of Medical Engineering & Technology”. pp. 1-12, 2007.

[53] Haykin, S. Neural Networks: “A comprensive foundation”. New Jersey - USA: Prentice Hall, 1999.

[54] Martín del Brío B., Sanz A. Redes neuronales y sistemas borrosos. Madrid: Alfaomega, 2007.

[55] Flórez R., Fernández J.M. Las redes neuronales artificiales. Editorial Netbiblo, 2008. 152 p.

[56] S. E. Fahlman and C. Lebiere. The cascade-correlation learning architecture. Technical Report CMU-CS-90-100, School of Computer Science, Carnegie Mellon University, August 1991.

[57] Jim C. Bezdek. Fuzzy Mathematics in Pattern Classification. PhD thesis, Applied Math. Center, Cornell University, Ithaca, 1973.

[58] Jang, J.-S. R, Sun C. -T, Mizutani E,Neuro - Fuzzy and Soft Computing: a computational approach to learning and machine intelligence, Prentice Hall, 1997.

[59] Timothy J. Ross, Fuzzy Logic with Engineering Applications, John Wiley & Sons, 2004.

[60] Hanley James A, McNeil Barbara J, The Meaning and Use of the Area under a Receiver Operating Characteristics (ROC) Curve. Departament of Epidemiology and Health, McHill University, Montereal, Canadá (J.A.H.) and the Departament of Radiology, Harvard Medical School and Brigham and Women's Hospital, Boston, MA (B.J.M.) Revision Receiver Dic. 15 de 1981.

[61] A.R. van Erkel, P.M. Th. Pattynama, Receiver operating characteristic (ROC) analysis: Basic principles and applications in radiology, European Journal of Radiology 27, pp. 88-94,1998.

[62] E.Y.-K. Ng, A review of thermography as promisin non-invasive detection modality for breast tumor, International Journal of Thermal Sciences, 849-859, 2009.

[63] E. Y. K Ng, L. N. Ung, F. C. Ng, and L. S. J. Sim, "Statistical analysis of healthy and malignant breast thermography", Journal of Medical Engineering & Technology, vol. 25, No. 6, pp. 253-263, Nov./Dec. 2001.

[64] N. Scales, C. Herry, and M. Frize, "Automated image segmentation for breast analysis using infrared images", in Proc. 2004 IEEE EMBS, pp. 1737-1740.

[65] X. Tang, and H. Ding, "Asymmetry analysis of breast thermograms with morphological image segmentation", in Proc. 2005 IEEE EMBS, pp. 1680-1683.

[66] C. Olson, "Constrained hough transforms for curve detection" [online], in Computer Vision and Image Understanding, vol. 73, No. 3, pp. 329-345, Marzo 1999 [citado: Junio 16 de 2009]. Disponible desde: http://faculty.washington.edu/cfolson/papers/pdf/cviu99.pdf

[67] Wiecek, M. Strzelecki, T. Jakubowska, M. Wysocki,Drews-Peszynski, Advanced Thermal Image Processing,The Biomedical Engineering Handbook, 3ed, CRC Press, 2006.

[68] Hairong Qi, Phani Teja Kuruganti, Wesley E. Snyder, Detecting Breast Cancer from Thermal Infrared Images by Asymmetry Analysis,he Biomedical Engineering Handbook, 3ed, CRC Press, 2006.

[69] R.G. Dumitrescu, I. Cotaria, Understanding breast cáncer risk – where do we stand in 2005?, J. Cell. Mol. Med. Vol 9, No 1, 2005. pp. 208-221.

[70] Ron Kohavi, George H. John, Wrappers for feature subset selection, Artificial Intelligence, Elsevier Science B.V., 1997. pp 273-324.

[71] THE MATHWORKS. Bioinformatics Toolbox, User's Guide. Natick: The Mathworks Inc. 2007.

[72] THE MATHWORKS. Neural Network Toolbox, User's Guide. Natick: The Mathworks Inc. 2007.

[73] D. F. Specht, Probabilistic neural networks for classification, mapping,or associative memory, in Proc. IEEE Int. Conf. Neural Networks, July1988, pp. 525–532

Universidad Industrial de Santander

ANEXO A. CONSENTIMIENDO INFORMADO

Estimada Paciente:

Las enfermedades de los senos son cada vez mas frecuentes en las mujeres en todo el mundo. Por esto, los investigadores del proyecto de investigación "**EVALUACIÓN DE LA TERMOGRAFÍA INFRARROJA EN LA DETECCIÓN DEL CARCINOMA DE GLÁNDULA MAMARIA**" realizado por miembros del Departamento de Patología de la Universidad Industrial de Santander (UIS) y de la Unidad de Oncología del Hospital Universitario (HUS) de Santander, en unión con el Grupo de Investigación de Conectividad y Procesado de Señal (CPS) y el grupo de investigación en Innovación y Desarrollo Tecnológico de Unisangil (IDENTUS) preocupados ésta situación estamos realizando este estudio que tiene como objetivo conocer la temperatura de las distintas partes del seno y sobre todo la temperatura específica del trastorno que pudiera tener el seno enfermo. Esto permite saber si éste método es útil para detectar el cáncer de seno en mujeres de cualquier edad.

La técnica de termografía infrarroja, es un examen que se puede realizar en cualquier parte del cuerpo, pero en éste estudio lo vamos a realizar a los senos únicamente. Se hace mediante el uso de una cámara diseñada especialmente para esto. La cámara se utiliza como una cámara fotográfica, de manera que registra la temperatura superficial de los senos. Lo que se observa después de la toma es una imagen de los senos con distintos colores en donde las zonas más calientes se ven de un color diferente al de las zonas más frías. Eso hace que si en el seno hay alguna masa, ésta pudiera resultar más fría o más caliente que el resto del seno y haría pensar que existe enfermedad en ese sitio.

La toma de la imagen es únicamente de los senos y no de su rostro ni do otro sitio del cuerpo. La cámara no estará en contacto físico con usted, no genera rayos x y es un procedimiento rápido e indoloro.

Este estudio se realizará en 200 pacientes, que es el número de personas necesarias para poder hacer un análisis valedero de los resultados. Su participación es completamente voluntaria y la realización de la técnica no implica ningún costo para usted como tampoco habrá compensación económica alguna. Los miembros del grupo investigador estarán en disposición de brindarle ahora y en el futuro cualquier información o pregunta que le surja acerca de los resultados o del procedimiento; para eso se suministrará en el momento de su valoración los datos de los responsables de aclarar dudas al respecto.

La información generada por este estudio es estrictamente confidencial y se mantendrá su privacidad. La información del estudio no será utilizada para generar beneficios

económicos. Usted es libre de rehusar a participar en este estudio en cualquier momento sin que esto conlleve a cambios en su futuro cuidado.

Usted tiene derecho a conocer los resultados de los estudios realizados cuando lo desee, una vez se haya realizado el análisis de las imágenes así como de solicitar que no sean incluidos en las conclusiones del trabajo. De igual forma el grupo investigador podrá tomar la decisión de retirarla del estudio si lo considera conveniente.

PARTICIPANTE

Yo _____ Firma _____

Cédula No. _____ de _____

He leído y recibido copia del presente consentimiento informado. Habiendo comprendido el significado de la investigación declaro estar debidamente informada y consiento en participar en este el estudio.

Ciudad: _____ Fecha: _____Hora:_____

TESTIGOS

Nombre _____ Firma _____

Nombre:_____ Firma _____

Ciudad: _____ Fecha: _____ Hora:_____

INVESTIGADOR QUE BRINDA EL CONSENTIMIENTO

Nombre:_____ Firma _____

Telefóno: _____

Ciudad: _____ Fecha: _____ Hora:_____

ANEXO B.FORMATOS DE RECOLECIÓN DE LA INFORMACIÓN

Universidad Industrial de Santander

ENCUESTA SOBRE LOS FACTORES ASOCIADOS SOCIODEMOGRÁFICOS, HEREDITARIOS Y HORMONALES ENDÓGENOS EN LAS PACIENTES CON CARCINOMA INFILTRANTE DE LA GLÁNDULA MAMARIA EN EL DEPARTAMENTO DE SANTANDER

Se guardará la confidencialidad de los datos y en ningún momento se revelará la identificación de los pacientes.

NOTA: Es importante que al registrar los datos tenga en cuenta las unidades de medida, marque con una X la respuesta en la casilla que corresponda y evite dejar espacios en blanco.

Nombre (s) y Apellidos:
Dirección: Barrio:
Tél. Fecha hoy: dd/mm/aa (/ /)

Código de las Imágenes: Frontal:_____
Lateral Der _____ Lateral Izq _____
Oblicua Der._____ Oblicua Izq_____
No. de Historia Clínica: Cédula:
Talla del Brasier: □ □ □

DATOS SOCIODEMOGRAFICOS

4. ¿Cuántos años cumplidos tiene?: □□ años
5. ¿Cuál es su fecha de nacimiento? (/ /) dd/mm/aa
6. En el último año, ¿dónde ha residido? Nombre del lugar (municipio y/o vereda)_____

Área urbana□ Área rural □

7. De acuerdo a su recibo de luz, ¿A cual estrato corresponde su vivienda? 1□ 2□ 3□ 4□ 5□ 6□
8. En su opinión, ¿a cuál de las siguientes razas pertenece usted? 1. Blanca □ 2. Mestiza □ 3. Negra □ 4. No sabe □ 5. No responde □
9. ¿Cuál es su Estado Civil?:

Soltero	□
Casado	□
Divorciado	□
Unión Libre	□
Otro	□

10. ¿Convive con pareja estable y permanente? Si □ No □
11. ¿Cuál fue el último grado de estudios que usted aprobó?

Tipo de enseñanza	No. de años
Ninguna	0□
Primaria	1□ 2□ 3□ 4□ 5□

Secundaria	6 · 7 · 8 · 9 · 10 · 11 ·
Técnica o Universitaria, incluyendo Postgrados	1 · ·2 · · 3 · · 4 · ·5 · ·6 · ·7 · ·8 · ·9 · ·10 ·

12. ¿A qué se dedicó la mayor parte del tiempo en el ultimo año?

Trabajó	·
Trabajó y estudió	·
Estudió (a)	·
Actividades del hogar	·
Buscó trabajo	·
Pensionado (a)	·
Retirado sin pensión	·
Otra	· ¿Cuál?

13. ¿En cuál de los siguientes rangos está el ingreso mensual de su familia (personas que aportan económicamente para el sostenimiento de su hogar)?

1. $0-$496.800 · 2. $496.800-$993.600 · 3. $993.600-$1987.200 · 4. $1987.200-$3.974.400 · 5. $3.974.400 o más · 6.No sabe · 7. Rehúsa contestar ·

COBERTURA Y ACCESO A LA ATENCIÓN MÉDICA

14. ¿A su familia alguna vez le aplicaron la encuesta del SISBEN?

Si · No · · ·Si “No” pase a la pregunta 14 No sabe/No recuerda · · ·Si “No” pase a la pregunta 14

15. Después del año 2001, ¿Le han aplicado la encuesta del SISBEN a su familia?

Si · No · No sabe/No recuerda ·

16. ¿En que nivel del SISBEN está clasificado? ·

17. ¿En el ultimo año ha estado o estuvo asegurado o afiliado a un plan de salud como cotizante?

Si · No · Si, pero no siempre · No sabe/No recuerda ·

18. ¿Actualmente esta asegurado o afiliado a un plan de salud como cotizante?

Si · No · No sabe/No recuerda ·

19. ¿En el ultimo año ha estado o estuvo asegurado o afiliado a un plan de salud como beneficiario?

Si · No · Si, pero no siempre · No sabe/No recuerda ·

20. ¿Actualmente esta asegurado o afiliado a un plan de salud como beneficiario?

Si · No · No sabe/No recuerda ·

21. ¿Actualmente a que entidad de salud está afiliado o es beneficiario?

1.Nueva EPS (ISS)	·	6. Fuerzas Militares, Policía Nacional	·
2.Administradora de régimen Subsidiado (ARS)	·	7. ECOPETROL	·
3.Empresa promotora de	·	8. Magisterio	·

Salud (EPS)

4.Empresa de Medicina prepagada ・ 9. Ninguna ・

5.Empresa Solidaria ・ 10. Otra, ¿Cuál?

22. Nombres de la entidad de salud a la cual está afiliado

_____

HISTORIA DE SALUD FAMILIAR

23. ¿Su mamá biológica tiene o tuvo cáncer?

Si ・ No ・ ・・Si “No” pase a la pregunta 23 No sabe ・ ・ Si “No sabe” Pase a la pregunta 23

24. ¿Qué edad tenía su mamá cuando le diagnosticaron cáncer?

・・ años

25. ¿En qué sitio/órgano su mamá tiene o tuvo cáncer?

_____

26. ¿Su mamá biológica está viva?

Si ・ ・・・・No ・ No sabe ・

27. ¿Su papá biológico tiene o tuvo cáncer?

Si ・No ・ ・・Si “No” pase a la pregunta 27 No sabe ・ ・Si “No” pase a la pregunta 27

28. ¿Qué edad tenía su papá cuando le diagnosticaron cáncer?

・・ años

29. ¿En qué sitio/órgano su papá tiene o tuvo cáncer?

_____

30. ¿Cuántas hermanas tiene usted? ・・ Si no tiene hermanas pase a la pregunta 31

31. ¿Alguna de sus hermanas tiene o tuvo cáncer de mama?

Si ・ No ・ ・・Si “No” pase a la pregunta 31 No sabe ・

32. ¿Cuántas hermanas tuvieron cáncer de mama? ・・・・

33. ¿Qué edad tenía(n) su(s) hermana(s) cuando le diagnosticaron cáncer de mama?

1. ・・ años 2. ・・ años 3. ・・ años

34. ¿Algunas de sus hijas tiene o tuvo cáncer de mamá?

Si ・ No ・ ・・Si “No” pase a la pregunta 34 No sabe ・

35. ¿Cuántas hijas tienen o tuvieron cáncer de mama? ・・

36. ¿Qué edad tenía(n) su(s) hija(s) cuando le(s) diagnosticaron cáncer de mama?

1. ・・ años 2. ・・ años 3. ・・ años

37. ¿Alguna de sus tías o primas tiene o tuvo cáncer de mama?

Si ・ No ・ ・Si “No” pase a la pregunta 39 No sabe ・

38. ¿Cuántas tías tienen o tuvieron cáncer de mama? ・・

39. ¿Qué edad tenía(n) su(s) tía(s) cuando le(s) diagnosticaron cáncer de mama?

1. ・・ años 2. ・・ años 3. ・・ años

40. ¿Cuántas primas tienen o tuvieron cáncer de mama?

・・・・・・Ninguna ・・Si “Ninguna” pase a la pregunta 39

41. ¿Qué edad tenía(n) su(s) prima(s) cuando le(s) diagnosticaron cáncer de mama?

1. ・・ años 2. ・・ años 3. ・・ años

ANTECEDENTES PERSONALES

42. ¿Qué edad tenía usted cuando tuvo su primera menstruación, regla o periodo?

・・・ ・・ años

43. ¿Ha tenido embarazos?

Si ・ No ・ ・Si “No” pase a la pregunta 45 ¿Cuántos? ・・

44. ¿Ha tenido recién nacidos muertos?
Si · No · Si la respuesta es “si” ¿cuantos?_____
45. ¿Ha tenido abortos?
Si · No · Si la respuesta es “si” cuantos abortos?_____

46. ¿A que edad tuvo su primer embarazo?
·· años

47. ¿Alimentó con leche materna a su(s) hijo(s)?
Si · No ·
¿Cuánto tiempo alimento con leche materna a su(s) hijo(s)?
1. Hijo_____(meses)
2. Hijo_____(meses)
3. Hijo_____(meses)
4. Hijo_____(meses)
5. Hijo_____(meses)
6. Mas? Si · No · Si la respuesta es “si”, sumé el tiempo (aproximadamente) de los restantes en que los amamanto_____

48. ¿Ha sido diagnosticada alguna vez alguno de estos canceres?:
Mama Si · No ·
Ovario Si · No ·
Útero Si · No ·

49. ¿Ha transcurrido más de 12 meses desde su última menstruación?
Si · No ·
50. ¿Cuál fue la fecha de su última regla (el primer día de sangrado o menstruación)?
(/ /) dd/mm/aa
51. ¿Recibió tratamiento hormonal para la menopausia?
Si · No ·
Si su respuesta es “Si”, ¿Durante cuanto tiempo los uso o los ha venido usando?_____
Si recuerda, qué tipo de medicamento ¿usó?_____

52. ¿Recibe tratamiento hormonal para la menopausia?
Si · No ·
Si su respuesta es “Si”, ¿Durante cuanto tiempo los ha venido usando?_____
Si recuerda, qué tipo de medicamento ¿usó?_____

53. ¿Le han tomado mamografías en los últimos 3 años?
Si · No ·
Si su respuesta es “Si”, ¿la última mamografía fue tomada?
En el último año ···· Entre 1 y 2 años ···· Entre 2 y 3 años ·

54. ¿Alguna vez le han practicado cirugía o biopsia en los senos?
Si · No · · Si “No” pase a la pregunta 54

55. ¿Hace cuanto tiempo le realizaron la última intervención (cirugía o biopsia)?
En los últimos 3 meses ・・ Entre 3 y 12 meses ・・・・Entre 1 y 2 años ・・・・
¿Más años?¿Cuantos? _____ ・

56. ¿En que seno le realizaron el procedimiento (cirugía o biopsia)?
Derecho ・ Izquierdo ・

57. ¿Ha recibido alguna vez tratamiento de radioterapia?

Si ・ No ・ Si la respuesta es “si”, por cuanto tiempo?_____

INFORMACION DEL ESTADO DURANTE LA TOMA DE LA TERMOGRAFIA

58. ¿Ha realizado alguna actividad (caminata, bronceado, etc.) donde se haya expuesto prolongadamente al sol durante los últimos 5 días?
Si ・ No ・
Si la respuesta es “si” ¿Por cuánto tiempo? _____horas

59. ¿Hoy usó lociones, cremas, polvos o algún tipo de maquillaje en el área?
Si ・ No ・ ¿Cuál? _____
60. ¿Hoy usó desodorante o antitranspirante?
Si ・ No ・
61. ¿Realizó alguna terapia física en las últimas 24 horas?
Si ・ No ・
62. ¿Realizó algún ejercicio físico 4 horas antes del examen?
Si ・ No ・
63. ¿Ha tomado algún medicamento para el dolor o vaso dilatador el día del examen?
Si ・ No ・

REGISTRO DE LA INFORMACIÓN CLÍNICA

Se guardará la confidencialidad de los datos y en ningún momento se revelará la identificación de los pacientes.

NOTA: Es importante que al registrar los datos tenga en cuenta las unidades de medida, marque con una X la respuesta en la casilla que corresponda y evite dejar espacios en blanco.

Nombre (s) y Apellidos:
Dirección: Barrio:
Tél: Fecha hoy dd/mm/aa
(/ /)

Código de las Imágenes:Frontal:_____Lateral Der_____ Lateral Izq _____ Oblicua Der._____Oblicua Izq_____
No. de Historia Clínica: Cédula:
Estatura: _____ Peso:_____ IMC:_____

DATOS DEL EXÁMEN CLÍNICO

1. ¿La paciente presenta dolor en las glándulas mamarias?
   Si□ No □ Si la respuesta es "si", ¿En cuál glándula mamaria?
   Derecha □ Izquierda.
   En cuál región de la glándula mamaria?
   CSE □ CSI □ CIE □ CII □ Centro □
2. ¿La paciente se ha palpado masa(s) en la(s) glándula(s) mamaria(s)?
   Si□ No □ Si la respuesta es "si", ¿En cuál glándula mamaria?
   Derecha □ Izquierda
   En cuál región de la glándula mamaria?
   CSE □ CSI □ CIE □ CII □ Centro □
3. La paciente presenta asimetría en la inspección de las glándulas mamarias?
   Si□ No □
4. ¿La paciente presenta Telorrea?
   Si□ No □ Si la respuesta es "si" ¿La Telorrea es espontánea? Si□ No □,
   Si hay telorrea, tipo de secreción: Sanguinolenta □ Verdosa□ Blanquecina □
   Serosa □
5. ¿La paciente presenta masa(s) con dimensiones superiores a 2 *cm* en la glándula mamaria?
   Si□ Posiblemente □ No □ Si la respuesta es "si" o "Posiblemente",
   ¿En que región se ubica(n) la(s) masa(s)?
   CSE □ CSI □ CIE □ CII □ Centro □
   ¿Las masa(s) es (son)?
   Móvil(es) □ Fija(s) □
6. ¿La paciente presenta alteraciones en la piel de la glándula mamaria?
   Si□ No □ Si la respuesta es "si",
   ¿Qué tipo de anormalidad existe?

Eritema · Edema(piel de naranja) · Nodulaciones · Retracción · Ulceración ·

¿En qué región presenta la anormalidad de la piel?

CSE · CSI · CIE · CII · Pezón y areola ·

7. ¿La paciente presenta anomalías en los nódulos linfoides axilares?

Si · No · Si la respuesta es “si”,

¿Qué nivel de anomalía se encontró?

N1(menores a 2 *cm*) · N2 (mayores a 2 *cm*) · N3 ·

8. ¿La paciente presenta nódulo (s) supraclavicular(es)?

Si · No · Si la respuesta es “si”,

¿Qué nivel de anomalía se encontró?

N1(menores a 2 *cm*) · N2 (mayores a 2 *cm*) · N3 ·

9. Tiene resultado de mamografía?

Si · No · Si la respuesta es “si”,

Cuál es el resultado de la mamografía?

BIRADS 0 · BIRADS I · BIRADS II · BIRADS III · BIRADS IV · BIRADS V ·

ANEXO C. ESPECIFICACIONES DE LA CÁMARA INFRARROJA FLUKE TI50

		Fluke Ti55	Fluke Ti50
Óptica	**Thermal**		
	Campo de visión*	23° horizontal x 17° vertical	
	Campo de visión instantáneo*	1,30 mrad	
	Distancia focal mínima*	0,15 m	
	Sensibilidad térmica (NETD)	≤0,05 °C a 30 °C	≤0,07 °C a 30 °C
	Adquisición de datos/frecuencia de imagen del detector	60 Hz/30 Hz	
	Enfoque	SmartFocus; enfoque continuo con un dedo	
	Zoom digital de la image infrarroja	2x, 4x, 8x	2x
	Tipo de detector	Matriz de plano focal de 320 x 240 de óxido de vanadio (V0x) con microbolómetros no refrigerados y paso de 25 micrones	
	Banda espectral	De 8 µm a 14 µm	
	Mejora de la imagen digital	Realce automático permanente	
	Visualización (sólo modelos Fusion)		
	Modos de funcionamiento en pantalla	Imagen totalmente infrarroja. Imagen totalmente visible. Fundido de imágenes visible y térmica. Imagen en imagen	
	Cámara de luz visible	1280 x 1024 píxeles a todo color	
	Zoom digital de luz visible	2x, 4x, 8x	2x
Medida de la temperatura	Rango calibrado de temperatura	De -20 °C a 600 °C en 3 rangos	De -20 °C a 350 °C en 2 rangos
		Rango 1 = de -20 °C a 100 °C	Rango 1 = de -20 °C a 100 °C
		Rango 2 = de -20 °C a 350 °C	Rango 2 = -de 20 °C a 350 °C
		Rango 3 = de 250 °C a 600 °C	-
	Precisión	± 2 °C o 2% (la mayor de ambas)	
	Modos de medida	Punto central, zona central (mínimo, máximo y promedio de área), puntos/zonas desplazables, anotaciones de campo/texto definidas por el usuario, isotermas, detección automática de puntos fríos y calientes, alarma visible de temperaturas por encima y por debajo del valor establecido	Punto central, zona central (mínimo, máximo y promedio de área)
	Corrección de emisividad	De 0,1 a 1,0 (en incrementos de 0,01)	
Presentación de la imagen	Pantalla digital	Pantalla digital de 5 pulg. de alta resolución	
	Retroiluminación de pantalla	Pantalla LCD a color visible a la luz del sol	
	Salida de vídeo	Vídeo compuesto RS170 EIA/NTSC o CCIR/PAL	
	Paletas de color	Escala de grises, escala de grises inversa, rojo y azul, alto contraste, metal caliente, hierro (Ironbow), ámbar, ámbar inversa	
Lentes opcionales	**Lente teleobjetivo de 54 mm**	Lente de germanio de alta precisión	
	Campo de visión	9° horizontal x 6° vertical	
	Campo de visión instantáneo	0,47 mrad	
	Distancia focal mínima	0,6 m	
	Lente de ángulo amplio de 10,5 mm	Lente de germanio de alta precisión	
	Campo de visión	42° horizontal x 32° vertical	
	Campo de visión instantáneo	2,45 mrad	
	Distancia focal mínima	0,3 m	
Almacenamiento de imágenes y datos	Soporte de almacenamiento	Tarjeta CompactFlash con capacidad para almacenar más de 1.000 IR imágenes (tarjeta estándar de 512 MB)	
	Formatos de archivo compatibles	Archivo radiométrico con datos de 14 bits. Exportable a JPEG, BMP, PCX, PNG, PSD.	
Interfaces y software	Interfaz	Lector de tarjeta CompactFlash incluido	
	Software	SmartView; software completo de análisis y realización de informes incluido	
Láser (sólo en modelos IR-Fusion)	Clasificación	Clase II	
	Puntero láser	Punto láser visible en la pantalla al combinar imágenes visibles y térmicas	
Controles y ajustes	Controles de configuración	Fecha/hora, unidades de temperatura en C/F, idioma, escala, intensidad de pantalla LCD (alta/normal/baja)	
	Controles de imagen	Nivel, rango, ajuste automático (continuo/manual)	
	Indicadores en pantalla	Estado de la batería, emisividad del objetivo, temperatura de fondo y reloj en tiempo real	
Alimentación	Tipo de batería	Batería inteligente de ión-litio, recargable de sencilla sustitución	
	Vida útil de la batería	Funcionamiento continuo durante 3 horas (2 horas en modelos con IR-Fusion)	
	Carga de batería	Cargador inteligente de 2 puertos con toma de red CA	
	Funcionamiento CA	Adaptador de CA de 110/220 V CA, 50/60 Hz	-
	Ahorro de energía	Modos "Apagado" y "En espera" automáticos (especificados por el usuario)	
Diseño ambiental y mecánico	Temperatura de trabajo	De -10 °C a +50 °C	
	Temperatura de almacenamiento	De -40 °C a +70 °C	
	Humedad relativa	Del 10% al 95% sin condensación, en funcionamiento y almacenamiento	
	Resistente al agua y al polvo	IP54	
	Peso (baterías incluidas)	1,95 kg	
	Tamaño de la cámara (LxAxF)	162 x 262 x 101 mm	
Otras especificaciones	Garantía	2 años	

ANEXO D. CÓDIGO EN MATLAB DE LA TRANSFORMADA DE HOUGH PARABÓLICA

```
function max_store_mod = hough_parabola(I)

%Parabola Hough Transform.

%% Inicio de Programa
f = double(I);

[M,N] = size(f);
h = 1:M;
nh = length(h);
k = 1:N;
nk = length(k);

t = linspace(-1.5, 0, 17);
t = [t -fliplr(t(1:end-1))];
nt = length(t);

pmin = 15;
pmax = 30;
p = linspace(pmin, pmax, pmax - pmin + 1);
np = length(p);

[x, y, val] = find(f);
x = x - 1; y = y - 1;

%% Creando espacio de parametros

h = zeros(nh, nk);
h3d = zeros(nh, nk, np);

%% Preparando matrices que se usaran en el kernel

x_matrix = repmat(x, 1, nt);
y_matrix = repmat(y, 1, nt);
val_matrix = repmat(val, 1, nt);
t_matrix = repmat(t, size(x_matrix, 1), 1);

%% Kernel

for j = 1:np

    h_matrix = x_matrix + p(j)*t_matrix.^2;
    k_matrix = y_matrix - 2*p(j)*t_matrix;

    h_bin_index = round(h_matrix + 1);
    k_bin_index = round(k_matrix + 1);
```

```
    neg = find(h_bin_index <= 0 | h_bin_index > nh);
    h_bin_index(neg) = 1;
    val_matrix(neg)  = 0;
    neg = find(k_bin_index <= 0 | k_bin_index > nk);
    k_bin_index(neg) = 1;
    val_matrix(neg)  = 0;

    h   = h + full( sparse(h_bin_index(:), k_bin_index(:), val_matrix(:), nh, nk) );

    h3d(:,:,j) = h3d(:,:,j) + h;
    [ih, jk] = find(h == max(h(:)));

    if j == 1
       max_store = [ih jk p(j)*ones(size(ih,1),1) h(ih(1),jk(1))*ones(size(ih,1),1)];
    else
       max_store = [max_store; ih jk p(j)*ones(size(ih,1),1) h(ih(1),jk(1))*ones(size(ih,1),1)];
    end

    h = zeros(nh, nk);
end

%% Estableciendo picos absolutos

max_store_mod = h_null_peaks(max_store, 1);
```

function P = parabola_construct(parabola, M, N)

```
h = parabola(1, 1);
k = parabola(1, 2);
p = parabola(1, 3);

a = -1/(4*p);
b = k/(2*p);
c = -k^2/(4*p) + h;
y = 1:N;
x = a*y.^2 + b*y + c;

se = strel('diamond', 2);
P = poly2mask(y, x, M, N);
P = imclose(P, se);
P = 1 - del_areas(1-P, 1);
```