# Validación Cuantitativa de Información Termográfica para el Pre-diagnóstico de Cáncer de Mama


TESIS DE MAESTRÍA

**María Yaneli Ameca Alducin**

**Institución**
**Laboratorio Nacional de Informática Avanzada**
**Maestría en Computación Aplicada**

**17 de Diciembre de 2012**

# Validación Cuantitativa de Información Termográfica para el Pre-diagnóstico de Cáncer de Mama

*Tesis para obtener el grado de Maestro*
**Maestría en Computación Aplicada**

*Dirigida por el Doctor*
**Dr. Efrén Mezura Montes**
**Dr. Nicandro Cruz Ramírez**

**Institución**
**Laboratorio Nacional de Informática Avanzada**
**Maestría en Computación Aplicada**

**17 de Diciembre de 2012**

# Agradecimientos

En primer lugar quiero agradecerle a Dios por estar conmigo en cada paso que doy. A mi hijo Santiago por impulsarme con tu amor a continuar superandome. A mis padres por su apoyo y cariño incondicional. A mis hermanos Benito, Galan, Christian, Omar y a mis hermanas Gina y Moyra por mantener la unidad familiar y estar a mi lado. A mis Asesores Dr. Efrén y Dr. Nicandro por se mis mentores y guías durante la realización de la tesis. Al Oncólogo Enrique Martín del Campo Mena por motivar este trabajo. A la Abuelita Julia de Santi por ser un gran apoyo. A mis compañeros y ahora Amigos Rocío, Lupita, Mayra, Josue, Alfredo, Felipe, Gabriel, Roberto, Aldo y Nancy. A mis amigos y amigas por darme aliento. A Conacyt por brindarme el apoyo económico para realizar la maestría y al proyecto de Conacyt 79809.

# Índice

# Índice de figuras

# Índice de Tablas

# Capítulo 1

# Introducción

## 1.1. Antecedentes

El cáncer de mama en México afecta a un gran número de mujeres, sin importar su nivel socioeconómico [1]. La mayoría de los casos de cáncer son detectados en etapas avanzadas de la enfermedad, reduciendo con ésto las expectativas de vida [2].

Existen diferentes técnicas para detectar la enfermedad, entre ellas se encuentran: la autoexploración, la mastografía, el ultrasonido, entre otras [3, 4, 5]. El estudio por excelencia para dar un diagnóstico de cáncer de mama es la mastografía [6], pero debido a que el cáncer de mama agrupa a diferentes variantes de la enfermedad, hay casos en donde dicho estudio no da un diagnóstico acertado, por ejemplo, en mujeres menores a 40 años en etapas tempranas de la enfermedad, la mastografía pudiera no llegar a detectar la enfermedad, debido a que la densidad del tejido mamario es mayor entre más joven sea la paciente [7].

Para coadyuvar en el proceso del diagnóstico del cáncer de mama, se cuenta en una clínica local con un estudio llamado Termografía, que consiste en detectar variaciones de temperatura en la piel de la mama, mediante una cámara de rayos infrarojos [8]. En él se detectan variaciones de temperatura en zonas donde pudiera existir un tumor, ya que en estas regiones se presenta un aumento de temperatura, debido al flujo de sangre que necesita el tumor para crecer [9].

En este estudio particular, el oncólogo es el encargado de interpretar y asignar valores a parámetros que definió en base a su experiencia, además considera datos adicionales como el historial clínico del paciente; de esta manera obtiene un pre-diagnóstico.

Existe un área dentro de la Inteligencia Artificial conocida como Aprendizaje automático de Máquina (Machine Learning, ML), donde se estudian y modelan en una computadora los procesos de aprendizaje en sus diversas modalidades [10]. Estos modelos, se convierten a su vez en programas de

computadora que deben de mejorar con la experiencia [11].

De entre las diversas tareas que resuelve ML, se encuentran los problemas de predicción, en la que un caso particular es la clasificación, donde se tienen una serie de datos compuestos por atributos, y cada tupla tiene asignada una clase. El objetivo es construir un modelo para predecir la clase de esa tupla dados los valores de sus atributos.

Las técnicas de clasificación, de manera general, funcionan como sigue: se tiene un conjunto de datos (que se dividen en 2 conjuntos, uno de entrenamiento y el otro de prueba), representados por una lista de características conocidos como atributos. El problema consiste en encontrar una función $f(\mathbf{x})$ llamada hipótesis que clasifique dichos ejemplos [12]. Existen diferentes técnicas para realizar clasificación, entre estas técnicas encontramos: K-NN, discriminante lineal, árboles de decisión, Naive Bayes, Redes Bayesianas, redes neuronales [10].

## 1.2. Planteamiento del Problema

Debido a que el estudio basado en termografía para el pre-diagnóstico del cáncer de mama es alternativo y que apenas se está comenzando a aplicar en México[13], no existe un sistema que apoye al pre-diagnóstico que da el oncólogo de acuerdo a su expertiz.

El oncólogo es el que interpreta los datos obtenidos por la cámara de Infrarojos y junto con el historial clínico que tiene de la paciente emite, a su juicio, un pre-diagnóstico.

Los datos que obtiene el oncólogo de la clínica local, no han sido tratados estadísticamente, además que el pre-diagnóstico puede estar altamente sesgado por su criterio (la observación personal puede interpretar de forma diferente la información de cada paciente).

Derivado de ello, la presente tesis pretende documentar el proceso de toma de estudio termográfico y el análisis de los datos obtenidos a partir de las imágenes térmicas. También se tratarán estadísticamente los datos, así como se documentará la aplicación de diferentes técnicas de clasificación a la base de datos del experto, en donde se presentan tanto casos positivos como negativos. El conjunto de datos cuenta con un número considerable de atributos(más de 15).

## 1.3. Objetivo general y específicos

### 1.3.1. Objetivo general

Analizar mediante técnicas estadísticas y de minería de datos el conjunto de datos obtenido del estudio termográfico para el pre-diagnóstico de cáncer de mama

### 1.3.2. Objetivos específicos

1. Documentar el método de la toma del estudio termográfico y su interpretación.

2. Establecer el estado del arte de la termografía mamaria

3. Aplicar estadística descriptiva del conjunto de datos

4. Aplicar técnicas de Correlación a las variables termográficas

5. Aplicar técnicas de clasificación al conjunto de datos

## 1.4. Hipótesis

Existe consistencia entre la información cuantitativa del conjunto de datos termográficos y el pre-diagnóstico dado por el oncólogo.

## 1.5. Justificación

La justificación del presente proyecto puede estar dada por, (1) con respecto al problema a resolver y (2) con respecto a las técnicas de clasificación.

1. Dada la novedosidad del pre-diagnóstico basado en termografía para el cáncer de mama en México y al éxito en los pre-diagnósticos obtenidos por el médico oncólogo en los casos que él reporta. Es pertinente realizar una validación cuantitativa entre la información termográfica y el pre-diagnóstico emitido por el médico oncológo para validar este tipo de pre-diagnóstico.

2. Derivado de una revisión preliminar de la literatura especializada, se tienen aplicaciones centradas en el análisis de las imágenes de termografía para la detección de cáncer mama [9, 14, 15, 16]. Sin embargo, se tiene reportado un escaso trabajo en aplicación de técnicas de clasificación sobre los datos obtenidos de la termografía y la forma en que realiza el pre-diagnóstico el médico oncólogo. La base de datos se presume compleja de clasificar debido a que el número de atributos es grande y a la posible presencia de ruido, atributos discretos y numéricos combinados, y posibles valores incompletos.

## 1.6. Alcances y Limitaciones

### 1.6.1. Alcances

- La presente tesis va a realizar la documentación del estudio termográfico realizado por el médico oncólogo.
- Se aplicarán diferentes clasificadores ya implementados al conjunto de datos termográficos.

### 1.6.2. Limitaciones

- La cantidad de casos obtenidos por parte del médico oncólogo pudiera limitar el desempeño de los clasificadores.
- Los casos están sesgados, debido a que en su mayoría los casos reportados son positivos.

## 1.7. Organización del documento de Tesis

- **Capítulo 2. Cáncer de mama**. Se define el cáncer y en específico el cáncer de mama y las diferentes variantes de ésta enfermedad, así como algunas técnicas utilizadas para su detección, las ventajas y desventajas de éstas.
- **Capítulo 3. Termografía**. Se define la termografía y se presenta su historia, así como el estado del arte en termografía y derivado de ello se desprende una clasificación del trabajo relacionado, para poder ubicar la contribución del presente trabajo.
- **Capítulo 4. Procedimiento termográfico**. Se describe el procedimiento de la toma del estudio termográfico, así como el análisis termográfico y obtención de variables termográficas.
- **Capítulo 5. Materiales y Métodos**. Se describe el conjunto de datos con el que se va a realizar el análisis. También se describen los métodos de estadística descriptiva que se aplicará al conjunto de datos. Así como la descripción de los clasificadores que se van aplicar a los datos termográficos.
- **Capítulo 6. Metodología y Resultados**. Se presentan los resultados obtenidos de la aplicación de estadística descriptiva y correlaciones a los datos termográficos, así como los resultados obtenidos de la aplicación de diferentes clasificadores.
- **Capítulo 7. Discusión**. Se discuten los resultados obtenidos en el capítulo previo.

- **Capítulo 8. Conclusiones y trabajos futuros**. Se presentan las conclusiones de esta tesis y los trabajos futuros a realizar.

# Capítulo 2

# Cáncer de Mama

## 2.1. Antecedentes

### 2.1.1. Estadísticas del cáncer de mama en México y sus proyecciones a futuro

Actualmente el cáncer de mama a nivel mundial ocupa la primera causa de muerte por cáncer en mujeres [17]. En los países subdesarrollados de América Latina el cáncer de mama y la mortalidad por esta enfermedad van en aumento, por el envejecimiento poblacional, los nuevos patrones reproductivos, el aumento en la exposición a factores de riesgo y el acceso limitado a una detección temprana de la enfermedad [18].

México actualmente es un país joven, pero esta población va a envejecer y con una mayor exposición a los factores de riesgo como son: exposición prolongada a estrógenos, vida sedentaria, alta ingesta de carbohidratos y poca fibra, la nuliparidad o reproducción tardía, antecedentes de cáncer de mama en la familia, se estima un aumento en el número de las personas enfermas [5].

Actualmente México ocupa el lugar 101 de incidencia de cáncer de mama y 135 de mortalidad, de 172 países (Agencia Internacional de Investigación en Cáncer IARC) [18], aunque ocupa un lugar menor en comparación a los países desarrollados; sin embargo si se considera la tendencia ascendente del envejecimiento de la población, podría llegar a alcanzar a los países desarollados en algunos años. En la Figura 2.1 [1] se muestra la mortalidad y el aumento de casos de cáncer de mama en México a partir del año 1957 a 2007, así como una proyección de la enfermedad a 12 años, en la gráfica se observa el aumento de casos y de mortalidad dado por el envejecimiento de la población.

En México el cáncer de mama había ocupado el segundo lugar como causa de muertes por tumores malignos en la mujer, pero a partir del año 2006, el cáncer de mama asciende al primer lugar [1], desbancando al cáncer cérvico

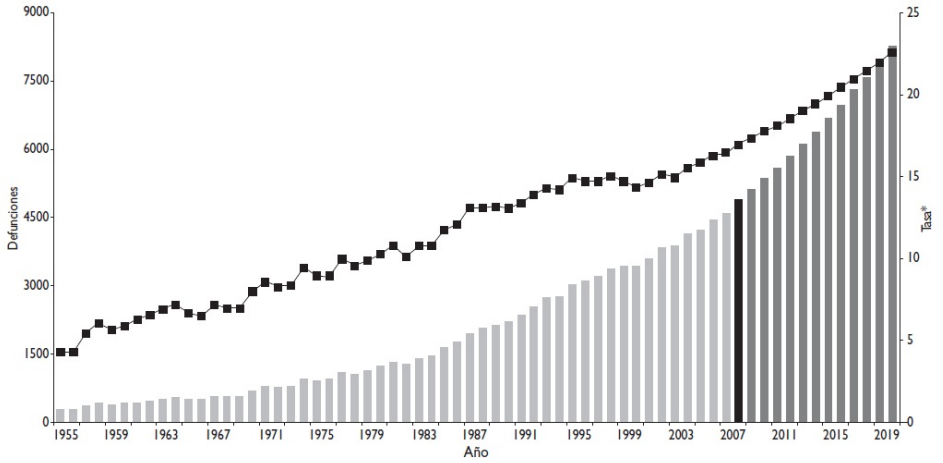


Figura 2.1: Tasa de casos de Cáncer de mama de 1955-2007 y proyecciones de 2008-2019 en México. *Tasa cruda por 100,000 mujeres de 25 y más años. Fuente: Bases de datos OMS/INEGI/SSA 1955-2007.

uterino, debido a la gran compaña de tamizaje realizada para la detección temprana de éste último.

Para realizar una campaña de detección como la del cáncer cérvico uterino, se necesitaría encontrar un tamizaje (evaluación masiva de sujetos asintomáticos respecto de una patología específica) para cáncer de mama que fuera barato, accesible e indoloro. Por desgracia, esta prueba no existe, la mastografía que es la prueba de pre-diagnóstico por excelencia, es realizada por un equipo muy costoso, no existe suficiente personal capacitado para realizar este estudio, y en comunidades alejadas es más difícil el acceso a estos recursos [19]. Cuando el paciente se presenta a realizarse un estudio por sospecha de cáncer de mama, usualmente es porque el tumor es evidente o la enfermedad está ya muy avanzada.

### 2.1.2. ¿Qué es el cáncer de mama?

Doll y Peto en 1986 definen al cáncer como una enfermedad en la que una o más células se alteran de tal manera que se multiplican de forma continua y producen millones de células con la misma alteración, algunas de las cuales se extienden a otras partes del cuerpo en incluso lo invaden [20]. Como se dividen y se multiplican las células, se forman masas de células que constituyen tumoraciones. Los tumores pueden ser benignos o malignos [21]. Los tumores benignos no contienen células cancerosas, no se extienden a otras partes del cuerpo y generalmente se pueden eliminar con cirugía, pero los tumores malignos contienen células anormales y se dividen sin control ni orden. Las células cancerosas pueden invadir y destruir el tejido de alrededor,

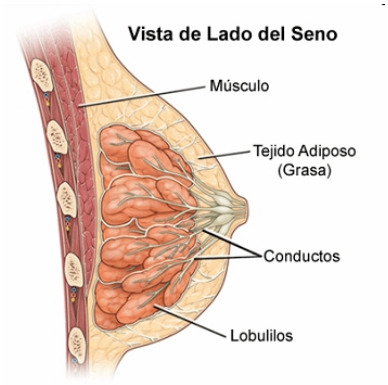


Figura 2.2: Principales estructuras de la mama.

también pueden separarse del tumor maligno y entrar al torrente sanguíneo o al sistema linfático, diseminándose a otras partes del cuerpo generando nuevos tumores; a ésto se le conoce como metástasis [22]. Por lo tanto el cáncer de mama (adenocarcinoma) es una enfermedad en donde la proliferación acelerada, desordenada y no controlada de las células pertenecientes a distintos tejidos de la glándula mamaria forman un tumor que invade los tejidos vecinos y puede haber metástasis a órganos distantes del cuerpo [23].

### 2.1.3. Glándula mamaria

La mama de una mujer adulta en promedio mide entre 10 y 12 cm. Su diámetro antero posterior es de entre 5 y 7 cm, se encuentra formada por tres estructuras principales: piel, tejido subcutáneo y glándula mamaria. En la Figura 2.2 [24] se muestran las estructuras de la glándula mamaria.

La función principal de la glándula mamaria es la producción y secreción de leche materna, producida por unas glándulas denominadas bulbos, que se agrupan para formar lobulillos y estos a su vez constituirán los lóbulos. Entre el tejido glandular se encuentra el estroma o tejido de sostén, formado por grasa, tejido conectivo, nervios, vasos sanguíneos y linfáticos. El tejido linfático constituyen el sistema de drenaje de la mama hasta los ganglios linfáticos de la axila, la zona clavicular y el tórax. El sistema linfático está formado por recipientes y vasos o conductos que contienen y conducen la linfa, que es un líquido incoloro formado por glóbulos blancos, en su mayoría linfocitos. Estas células reconocen cualquier sustancia extraña al organismo y liberan otras sustancias que destruyen al agente agresor. La mama pasa por diferentes etapas, dependiendo del estado hormonal de la mujer. En su mayoría el tejido mamario es glandular. En la menopausia el tejido glandular se atrofia, lo que justifica la mayor incidencia de cáncer de mama después de esta etapa [22].

### 2.1.4. Taxonomía de tipos de cáncer de mama

Los tipos de cáncer de mama generalmente son clasificados de la siguiente manera [25].

1. In situ
2. Carcinomas infiltrantes o invasivos.

Existen dos tipos de carcinoma in situ:

- *Carcinoma lobular in situ (CLIS)*: También llamado neoplasia (tumor) lobular. Se origina en los lóbulos o lobulillos de la mama (las glándulas fabricantes de leche). No atraviesa las paredes de éstos por lo que generalmente no se convierte en cáncer invasivo. No obstante, existen casos en los que sí puede desarrollarse y convertirse en un carcinoma lobular invasor.
- *Carcinoma intraductal o ductal in situ (CDIS)*: También se le llama carcinoma intraductal. Es el tipo más común de cáncer no invasivo de mama (16 %) [26], en el cual hay presencia de células anormales en el revestimiento de un conducto de la mama. En este caso las células cancerosas no se propagan a través de las paredes hacia el tejido adiposo del seno. El tratamiento incluye cirugía o radiación, que generalmente son favorables a la cura del padecimiento. No obstante, si no se tratan a tiempo pueden convertirse en invasivos.

Los carcinomas infiltrantes o invasivos son aquellos donde las células tienen la capacidad de atravesar la membrana basal, pudiendo así presentar metastasis en otras partes [26].

A continuación los tipos de cáncer de mama invasivos:

- *Carcinoma ductal infiltrante (o invasivo)*: Se origina en las glándulas productoras de leche. Puede extenderse hacia los canales linfáticos o a los vasos sanguíneos del seno y distribuirse a otras partes del cuerpo. Este es el tipo de tumor invasivo que más se presenta en los casos de cáncer de mama [26].
- *Carcinoma lobulillar infiltrante (o invasivo)*:Se origina en las glándulas productoras de leche y puede extenderse a otras partes del cuerpo. Se estima que entre 10 y 15 % de los cánceres invasivos pertenecen a esta clasificación. Es dificíl de detectar por examen físico o incluso en la mastografía, por su ubicación [23].
- *Carcinoma medular*: Se estima que es responsable del 5 % de todos los casos de cáncer de mama [23]. En él, las células cancerosas se encuentran agrupadas y en los bordes del tumor existen células del sistema

inmunitario que sirven para atacar y destruir las células anormales, así como a otros agentes extraños como bacterias o virus.

- *Carcinoma coloide*: Está formado por células que producen mucosidad. En términos médicos se le denomina carcinoma mucinoso. Pertenece al tipo de cáncer ductal invasivo y tiene un pronóstico favorable al tener menos probabilidades de propagación que el cáncer ductal invasivo o el lobular invasivo.

- *Carcinoma tubular*: El carcinoma tubular se caracteriza por la proliferación de glándulas o tubos bien diferenciados [23]. Existen menos probabilidades de que se propague fuera del seno, comparado con el cáncer ductal invasivo o el lobulillar invasivo. Es el responsable del 2 % de todos los casos de cáncer de mama.

- *Cáncer inflamatorio de mama*: No es muy común, representa apenas el 1 % de los casos de cáncer de mama, también denominado mastitis carcinomatosa. Es la forma más agresiva de cáncer de mama. Los síntomas son: la mama presenta un aspecto inflamatorio(piel enrojecida y caliente, con la apariencia de una cáscara de naranja). Las células cancerosas bloquean los vasos linfáticos de la piel [26], es decir que no se trata de una simple inflamación. Este tipo de cáncer tiene mayores probabilidades de propagación y su pronóstico es menos alentador que otros tipos.

## 2.2. Técnicas de Pre-diagnóstico de cáncer de mama

En la actualidad la mejor opción contra el cáncer de mama es una detección temprana del tumor, conllevando al aumento en las posibilidades de éxito del tratamiento [27]. En las siguientes secciones, se describirán brevemente las diferentes técnicas de pre-diagnóstico de cáncer de mama.

### 2.2.1. AutoExploración

La autoexploración es una técnica de detección del cáncer de mama, se basa en la observación y palpación que hace el paciente en sus mamas, permite detectar tumores más pequeños que los que pueda detectar el médico o la enfermera, pues el paciente estará más familiarizado con sus mamas y podrá detectar cualquier pequeño cambio. En las revisiones ginecológicas, el médico comprueba que no exista ninguna irregularidad en las mamas, también que no haya ninguna inflamación de los ganglios linfáticos axilares. La autoexploración es referida por algunos críticos como una herramienta de poca utilidad, ya que no detecta lesiones tempranas. Aunque en México

el 90 % de los casos son detectados por las pacientes, cuando detectan un abultamiento o nódulo, en estos casos ya se trata de un estado avanzado de la enfermedad [28]. En la Figura 2.3 [29] se da un ejemplo de la técnica de autoexploración.


Figura 2.3: Ejemplo de autoexploración mamaria. Fuente: A.D.A.M. Images.

### 2.2.2. Biopsia

Una vez detectado el tumor mediante una o varias de las técnicas mencionadas en el presente documento de tesis, se debe realizar una biopsia para confirmar el diagnóstico. La biopsia es la extirpación o extracción de tejido mamario con el fin de realizarle un examen patológico en busca de células cancerosas u otros trastornos [3, 30]. Una vez extraído el tejido mediante la biopsia, el patólogo examinará la muestra y determinará, si el tumor es benigno o maligno, en caso de ser maligno se puede determinar el estado del tumor, así como su capacidad para extenderse y la rapidez en la que lo hará. Se pueden realizar algunos tipos diferentes de biopsias como: biopsia estereotáctica de mama, biopsia de mama con ultrasonido y Tumorectomía. En la Figura 2.4 [13] se da un ejemplo de una biopsia.


Figura 2.4: Ejemplo de biopsia mamaria. Fuente: Malaga Vascular Institute.

### 2.2.3. Mastografía

Es una imagen plana de la glándula mamaria obtenida por rayos X de baja potencia para localizar zonas anormales en la mama. Esta técnica consiste en colocar la mama entre dos placas y presionarla durante unos segundos mientras se realizan las radiografías. No hay ningún peligro por las radiaciones de esta técnica, ya que son de baja potencia. Es una de las mejores técnicas para detectar el cáncer de mama en sus primeras fases. Una mastografía de escrutinio busca visualizar lesiones no palpables, aquellas menores a 0.5 cm si se trata de nódulos y calcificaciones nunca palpables por su tamaño, asimetrías en la densidad mamaria, y/o distorsión de la arquitectura de la glándula mamaria. Este estudio, aunque es la técnica por excelencia utilizada como pre-diagnóstico, tiene sus inconvenientes. Es difícil detectar irregularidades en mamas jóvenes, debido a que el tejido es denso; también se dificulta cuando se tienen implantes, si se tiene una sola mama, poder dar un pre-diagnóstico se ve afectado, porque para hacerlo se utiliza la comparación con la otra mama [5, 6, 28]. En la Figura 2.5 se da un ejemplo de una Mastografía.

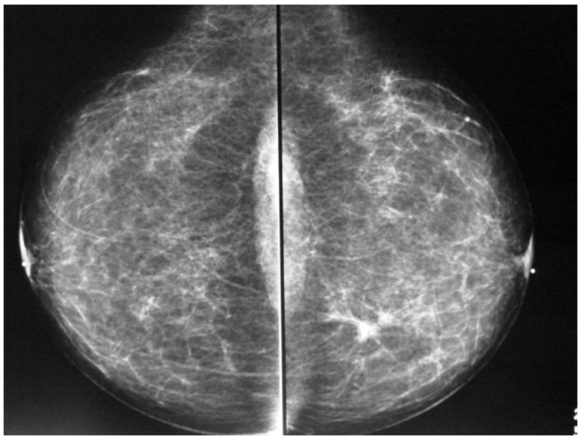

Figura 2.5: Ejemplo de mastografía mamaria. Fuente: imagen otorgada por el Cirujano Oncólogo Enrique Martin del Campo Mena.

### 2.2.4. Termografía

La termografía es una técnica alternativa para el pre-diagnóstico de cáncer de mama a través de imágenes infrarrojas. Es simple, no invasiva, barata, rápida, indolora e inofensiva. En ella se detectan variaciones en la temperatura de la piel durante un periodo de tiempo y utiliza imágenes térmicas para determinar la salud de la zona. Cada mama tiene un patrón térmico particular que no debe diferir en el tiempo, al igual que una huella digital. Las temperaturas de los pechos sanos y cancerosos son diferentes debido al metabolismo presente en el tejido con lesión. Los tumores, por hambre, aumentan la circulación de nutrientes a sus células mediante la apertura de los vasos sanguíneos existentes y creación de otros nuevos; a ésto se le conoce como angiogénesis. Estas actividades provocan un aumento de la temperatura en la superficie regional de la mama con lesión. La temperatura de la piel del seno con tumor puede ser de hasta 3.2°C mayor que la del tejido normal. Existen métodos para mejorar la detección de las diferencias de temperatura entre mamas sanas y enfermas. Éstos incluyen el enfriamiento de los senos con alcohol y la inmersión de las manos en agua fría antes del estudio, en un ambiente controlado. Al comparar los patrones térmicos tomados durante un periodo de tiempo con el patrón de línea de base normal, cualquier cambio significativo que se detecte es un indicador de que algo nuevo se está desarrollando dentro de la mama y que justifica una investigación. La detección temprana y precisa de los tumores pequeños de mama con estudios termográficos han sido reportados [8, 31, 7, 32]. En la figura 2.6 se presenta una visualización de las imágenes termográficas de las mamas.

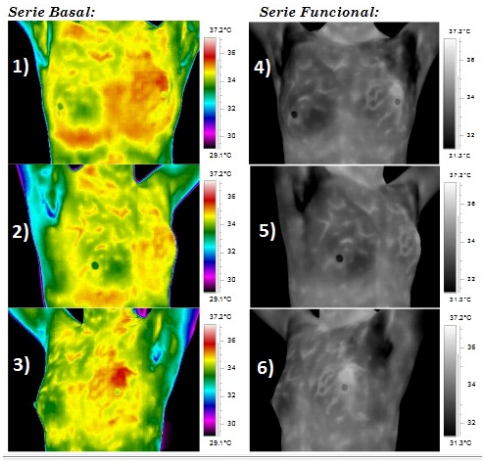


Figura 2.6: Ejemplo de imágenes obtenidas por un estudio termográfico. Fuente: imagen otorgada por el Cirujano Oncólogo Enrique Martin del Campo Mena.

### 2.2.5. Tomografía

La tomografía es una técnica de imagen que pone en evidencia los cambios del metabolismo glicolítico(glucosa), que de forma muy precoz se manifiestan a lo largo de todo el proceso tumoral. Su principal limitación en el cáncer de mama es la detección de lesiones tumorales de pequeño tamaño o de micrometástasis ganglionares axilares. Sin embargo ofrece una información muy importante en la estadificación de pacientes con alto riesgo, ante la sospecha de recidiva (aparición de la enfermedad) clínica o en la valoración de la respuesta terapéutica [33]. Existen diferentes tipos de Tomografía entre las que encontramos:

- *Tomografía axial computadorizada (TAC)*: Consiste en una técnica de rayos X, utiliza un haz giratorio, con la que se visualiza distintas áreas del cuerpo desde diferentes ángulos. Sirve para el diagnóstico de las metástasis, no del cáncer de mama propiamente dicho.
- *Tomografía por emisión de positrones (PET)*: Consiste en inyectar un radio fármaco combinado con glucosa que será captado por las células cancerosas, de existir un cáncer, pues éstas consumen más glucosa. El radio fármaco hará que se localicen las zonas donde se encuentre el tumor.

### 2.2.6. Ultrasonido

El ultrasonido de mama es una técnica para examinar de forma más detallada las anormalidades en la mama, detectadas por un médico


Figura 2.7: Ejemplo de ultrasonido mamario. Fuente: imagen otorgada por el Cirujano Oncólogo Enrique Martin del Campo Mena.

durante un examen clínico o a través de la mastografía, es un estudio complementario. Los estudios han demostrado que utilizar el ultrasonido en combinación con la mastografía puede resultar en una detección más precoz y más frecuente de cáncer de mama. Es utilizado debido a su capacidad para poder diferenciar entre formaciones solidas y líquidas. Sin embargo en un buen número de lesiones malignas uno de los factores que influye para la precisión en el diagnóstico del ultrasonido mamario es la densidad de las mamas. En un estudio realizado a más de 3,000 mujeres se logro demostrar que el ultrasonido detecto en un 17 % más el número de casos con cáncer que el examen clínico. El ultrasonido tiene un mayor éxito en casos donde las mamas son densas a diferencia de la mastografía, debido a que en la mastografía el tejido denso y el canceroso se presentan de tonalidad blanca haciendo difícil su pre-diagnóstico, sin embargo en el ultrasonido el tejido denso se ve blanco y las posibles lesiones cancerosas se ven en tonalidad oscura. Hoy en día el ultrasonido también es utilizado en mujeres con implantes mamarios, porque la mastografía no siempre es útil para mostrar anormalidades, también es utilizada en mujeres embarazadas ya que estás no se pueden exponer a los rayos X [34]. En la Figura 2.7 se muestra un ejemplo de ultrasonido mamario.

### 2.2.7. Comparativo de Técnicas de pre-diagnóstico de Cáncer de mama

En la Tabla 2.1 se presenta un comparativo entre las diferentes técnicas utilizadas para dar un pre-diagnóstico de cáncer de mama.

Tabla 2.1: Comparativo de las diversas técnicas de pre-diagnóstico de cáncer de mama.

| **Técnicas** | **Ventajas** | **Desventajas** |
|---|---|---|
| Autoexploración | | |
| | Es indoloro | Se reliza una detección tardía de las lesiones |
| | Barato | |
| | No invasivo | Si no se conoce bien la técnica, no se pueden detectar lesiones |
| | El paciente al conocer sus mamas, puede identificar posibles lesiones | |
| Biopsia | | |
| | El resultado es muy confiable, por el análisis patológico | Es invasiva, dolorosa y costosa |
| | No es una técnica de pre-diagnóstico, es de diagnóstico | Normalmente se hacen estos estudios cuando la lesión es muy evidente y/o avanzada |
| Mastografía | | |
| | Localiza la lesión en la mama | Es invasiva, dolorosa, costosa |
| | Alto nivel de sensibilidad arriba del 85 % y especificidad | No se cuenta con suficiente equipo, ni personal calificado |
| | Prueba de pre-diagnóstico altamente aceptada en el área médica | Es difícil de detectar lesiones en mamas densas, con implantes o en personas con una sola mama |
| | Puede ver calcificaciones que posiblemente en futuro podrían volverse cáncer | No se puede realizar en mujeres embarazadas por la radiación |

| Técnicas | Ventajas | Desventajas |
|---|---|---|
| Termografía | | |
| | Es indolora, barata, no invasiva | No es aceptada en el área médica |
| | Puede detectar lesiones en mamas densas, con implantes o una sola mama | No tiene una alta sensibilidad (76 %) y tiene una muy baja especificidad (<50 %) |
| | Se le puede realizar a mujeres embarazadas | No da la ubicación exacta de la lesión |
| | Puede llegar a detectar lesiones hasta con 8 años de anticipación, por los cambios metabólicos que ocurren en la mama | |
| Tomografía | | |
| | Es indolora | No logra detectar lesiones muy pequeñas |
| | | Es muy costosa |
| | | Falta de equipos y personal capacitado para interpretar el estudio |
| Ultrasonido | | |
| | Es indolora, barata, no invasiva | Es un estudio complementario a la mastografía, ya que no puede ver calcificaciones que algunas de ellas posiblemente se vuelvan cáncer |
| | Puede detectar lesiones en mamas densas, con implantes o una sola mama | No puede detectar algunos tipos de canceres |
| | Se le puede realizar a mujeres embarazadas | No esta disponible en todas partes y requiere de personal calificado para su interpretación |

# Capítulo 3

# Termografía mamaria

## 3.1. ¿Qué es la termografía mamaria?

Es una técnica de pre-diagnóstico mediante imágenes térmicas de las mama, la cual puede ayudar en la detección temprana de cáncer de mama [35]. La imagen de infrarrojos (IR) de la mama, también conocida como termografía de mama, es un estudio no invasivo e indoloro, no expone al paciente a radiación ionizante, en la que se mide la respuesta fisiológica de la mama, con esto nos referimos específicamente a los cambios de temperatura, el estudio tiene un bajo costo y es realizado sin tener contacto con el cuerpo del paciente [36]. Este estudio se basa en la diferencia entre las temperaturas. Cuando una mama está enferma los tumores malignos o benignos necesitan más irrigación de sangre y esto conlleva a un aumento de temperatura con respecto a una mama sana [9].

Esta tecnología fue inicialmente diseñada en EU para visión nocturna [15], pero también tiene aplicaciones en la medicina. Su uso en el campo de la oncología médica radica en que en general los tumores tienen un aumento de temperatura y angiogénesis, así como una mayor tasa metabólica, lo que se traduce en un aumento de grados de temperatura en comparación con el tejido normal circundante. La detección de estos rayos infrarrojos "puntos calientez el aumento de temperatura, pueden ayudar a identificar y diagnosticar el cáncer. La Termografía ha estado en uso desde la década de los 60's. En 1970 se puso en marcha un estudio a nivel nacional en EU. Sin embargo se redujo la IR en las primeras etapas del proyecto debido a los estudios no satisfactorios, esto pudo deberse, a las dificultades técnicas o a la interpretación muy variable y subjetiva de las imágenes, las tasas altas de falsos positivos y falsos negativos y al no ayudar a la localización del área de la cirugía [31]. Aunque, hallazgos anormales en las IR pueden ser de ayuda para dife-

renciar entre tumores malignos y benignos. En 1980 fue aprobada por la FDA como una herramienta complementaria para el pre-diagnóstico de cáncer de mama [36]. Sin embargo, su aplicación fue limitada por la tecnología, equipos muy grandes y la falta de herramientas informáticas para el análisis. Desde entonces se han hecho importantes avances en la tecnología para digitalización de imágenes de alta resolución y análisis de imagen con software que integra técnicas de IA [37, 7, 32]. En el pasado los equipos de medición de infrarrojos sólo eran capaces de detectar la variación de 0.5 a 1.0 °C y algunos equipos necesitaban nitrógeno líquido además de tener contacto con el pecho de los pacientes para detectar la temperatura. Las cámaras termográficas actuales son capaces de detectar cambios de temperatura a partir de 0.08 °C en adelante y no requieren contacto con el paciente [7].

## 3.2. Estado del arte de la termografía

De la revisión de la literatura especializada en termografía mamaria, se dividieron estos trabajos en tres categorías: introductorios, basados en imágenes y basados en datos[36, 19, 9, 14, 38, 39, 37, 13].

Los trabajos introductorios destacan las potencialidades de la termografía mamaria como técnica alternativa en el pre-diagnóstico de cáncer de mama [36]. Por otra parte, hacen reseñas histórica de la termografía, desde sus comienzos hasta el estado actual de la termografía [31]. Además se compara el desempeño de la termografía mamaria con otras técnicas de pre-diágnóstico como la mastografía [31]. Existe un trabajo que muestra la ventaja de utilizar la termografía mamaria en lugares inaccesibles para otras técnicas de pre-diagnóstico [19]. Desafortunadamente, debido a que estos trabajos son introductorios al tema, les falta información sobre los datos utilizados en estos estudios, así como el tipo de análisis que se llevo acabo.

En los trabajos basados en imágenes, en lo que respecta a la adquisición de imágenes térmicas no existe un método estándar, hay trabajos que tratan el tema sobre la calibración de la cámara de infrarrojos para obtención de las imágenes, así como el protocolo que se debe de seguir para la adquisición de las imágenes [8, 40]. En otro trabajo se realizo un estudio a 50 pacientes, con el objetivo de analizar en las imágenes térmicas las variaciones cíclicas de vascularidad cuando se presenta el periodo menstrual y en ausencia de éste; del resultado de este estudio se sugieren que la toma de imágenes se realice entre los días del 12 al 21 después de la menstruación [16]. Además, en un trabajo se aborda

el cómo se podría llegar a detectar las emisiones de calor en la parte interior del cuerpo. Para ésto se utiliza un nuevo método basado en una analogía con los circuitos eléctricos, basado en la resistencia que presentan los circuitos eléctricos al paso de energía y como los tejidos también ofrecen resistencia al paso del calor y a la pérdida de calor por cada uno de los tejidos hasta llegar a la superficie, se hacen particiones de la imagen en rebanadas hasta llegar así a la fuente del calor [15]. También existe un estudio comparativo entre k- means y Fuzzy c-means, las cuales son técnicas para obtener clusters a partir de las imágenes termográficas para detectar lesiones en las mamas. Fuzzy c-means tuvo mejor desempeño al encontrar más clusters [9]. Otra técnica utilizada en el ánalisis de las imágenes térmicas es la dimensión de fractales, debido a que las lesiones malignas presentan formas irregulares, si se logran detectar lesiones malignas con esta técnica [14].

Las investigaciones de la termografía mamaria basadas en datos, presentan análisis estadísticos de los datos termográficos. Como el estudio termográfico realizado a 3768 pacientes, a partir del cual se obtuvieron 3 tipos de termografías mamarias (normal, dudosa y anormal). Descubrieron que la supervivencia de vida era igual para cualquiera de los 3 grupos [38]. Existen estudios estadísticos masivos donde sólo se miden el número de estudios termográficos anormales, sin indicar como se llevaron a cabo los análisis [41]. En un estudio se logro encontrar una correlación significativa con el tamaño del tumor, estado ganglionar histológico y el grado malignidad de las lesiones con cáncer, con la termografía mamaria [39]. Por otro lado, se llevo a cabo un estudio termográfico a 90 pacientes, a este estudio le fueron aplicadas reglas de Bayes. Los resultados fueron las entradas de una red neuronal que logro obtener un 74 % clasificados correctamente para positivos-verdaderos (sensibilidad), a este estudio le hace falta mencionar como le va con los falsos-positivos(especificidad) [42]. En otro trabajo se indica que un pre-diagnóstico depende de la interpretación correcta de la distribución de la temperatura de la piel, aunque este trabajo tiene más de 40 años da a conocer algunos factores que pueden afectar la temperatura de la piel [43].

Dentro de los trabajos basados en datos se encuentran los que emplean técnicas de IA, acontinuación se mencionan algunos. En un trabajo se utiliza un frame basado en redes neuronales. Una de las redes neuronales se basada en imágenes termográficas, ésta obtiene una sensibilidad de 68 % y una especificidad del 40 % y la otra red neuronal basada en datos fisiológicos tiene una sensibilidad de 48 % y una especificidad del 80 % [37]. Existen varios estudios de termografía mamaria con redes

neuronales [44, 7, 32], entre ellos existe un trabajo que se destaca por su alta sensibilidad con un 97 % y especificidad del 82 %, obtenida mediante un software llamado Breast Scan Breast Scan Sentinel (SBS) [7]. Por otro lado, se tiene reportado un trabajo que contradice la afirmación del trabajo anterior, indicando que SBS obtiene una sensibilidad de 70 % y especificidad del 33 % muy por debajo de lo que se reporta. En este trabajo se presenta un software para el pre-diagnóstico de cáncer de mama llamado NoTouch BreastScan, éste obtiene una sensibilidad del 89 % y especificidad del 42 % al combinar la mastografía con la termografía [32].

A partir de los datos se pueden generar modelos númericos, los cuales tratan de modelar las enfermedad de las mamas [45, 46, 47, 48]. Uno de los trabajos, generá un modelo de cuatro cuadrantes de una mama, posteriormente llevan a cabo el análisis del modelo mediante la comparación de patrones de temperatura. Simulando númericamente un patrón de temperatura y genera la simulación de estrés en frio. La simulación se obtiene mediante la variación de los parámetros que ayuden al investigador a realizar una evaluación cualitativa [47]. En otro trabajo utilizando el modelo anterior se realiza un estudio a 34 voluntario, el estrés en frio se llevo a cabo mediante la aplicación alcohol en el antebrazo, el estudio tarda 3 horas para la predicción de la temperatura sin estrés y de 36 horas para la de estrés frio. Computacionalmente el modelo puede resultar muy costoso [47]. Otro modelo numérico propuesto es uno tridimensional tomando una mama normal y una mama con presencia de un tumor maligno. El modelo se basa en la ecuación de biocalor y tien como ventaja con respecto a otros modelos, el hecho de que es una representación más cercana de la mama real [46].

De los datos de la termografía mamaria se pueden obtener variables y a partir de estás se forma un score. De los trabajos que utilizan un score para dar un pre-diagnóstico, se destaca uno basado en la diferencia de temperaturas entre las mamas, entre el tejido circundante y el aumento de vascularidad; a partir de estos valores se obtiene 5 variables que forman un score a partir del cual se emite un pre-diagnóstico [49]. En otro estudio basado en score son tomadas en cuenta más variables para emitir un pre-diagnóstico y en base a esto en el trabajo se pudo señalar y orientar la incisión inicial para la realización de una biopsia [13].

De la revisión de la literatura de la termografía mamaria, surgió la clasificación mencionada previamente y se puede ver en la figura 3.1. Nuestro trabajo se va a ubicar en lo que corresponde a la termografía con técnicas de Inteligencia Artificial o en los basados en score. En este trabajo se describe el procedimiento de la toma del estudio ter-

mográfico mamario, el análisis de las imágenes térmicas, la descripción del conjunto de datos derivada de la termografía, la metodología para realizar los análisis estadísticos, así como la aplicación de las diferentes técnicas de clasificación. Pero algunos artículos la cantidad de casos (termografías) era muy altos. En este trabajo sólo se cuenta con 98 casos, de los cuales la mayoría tiene cáncer de mama (77 casos) y muy pocos sin la enfermedad (21 casos), esto hace que nuestra muestra este muy sesgada hacia la enfermedad, lo que implica a tener una especificidad baja.

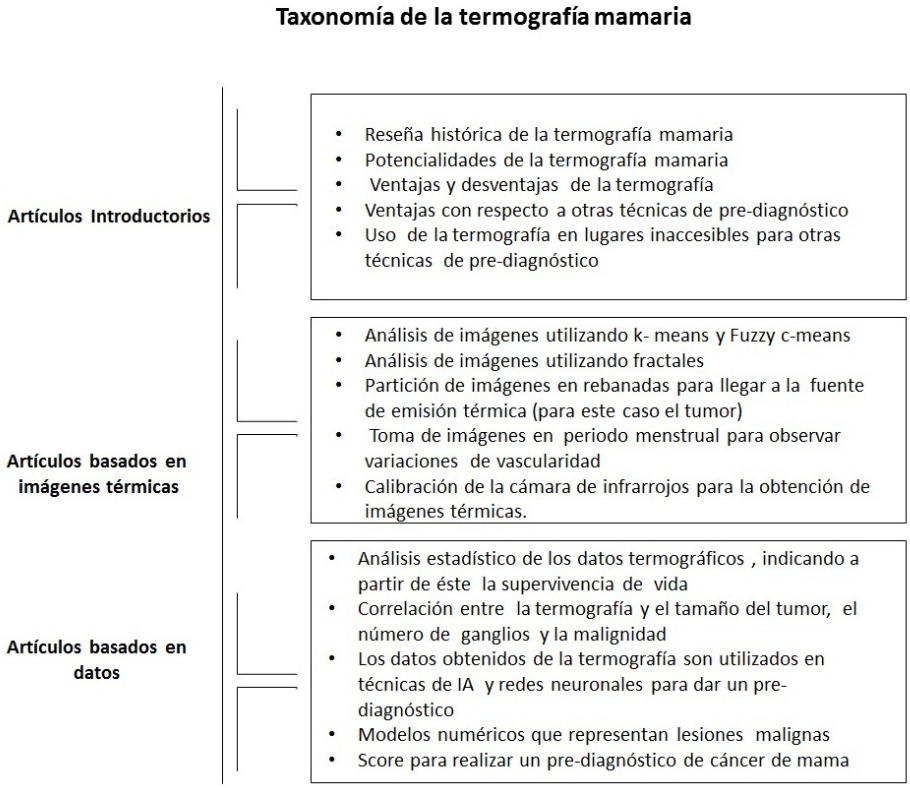


Figura 3.1: Taxonomía de la termografía mamaria.

# Capítulo 4

# Procedimiento Termográfico

En este capítulo se describirá el procedimiento del estudio termográfico mamario, realizado por el oncólogo experto. Este estudio consta de dos etapas: la toma de las imágenes termográficas de las mamas y la formación del score.

Para la toma del estudio al paciente, el tiempo requerido es de aproximadamente 15 minutos. Para la formación del score y darle formato al documento que es entregado al paciente se lleva aproximadamente 30 minutos por estudio. La persona que actualmente realiza la toma del estudio y la formación del score es un técnico entrenado por el médico oncólogo, la duración de su capacitación fue de 6 meses.

A continuación se describen las dos fases del estudio de la termografía mamaria.

## 4.1. Descripción de la toma de imágenes termográficas de las mamas

Para realizar el estudio de la termografía se utiliza una cámara FLIR A40 [50]. El estudio se realiza a una temperatura ambiente aproximada de 22 °C, a una distancia aproximada de un metro, dependiendo de la masa muscular de cada paciente. El paciente se descubre en su totalidad de la cintura hacia arriba, se calibra la cámara hasta que se visualice la forma de un triángulo en la imagen formado por la cabeza y los hombros como se puede observar en la figura 4.1, se deben observar en la imagen en su totalidad las 2 mamas, también es calibrada en el software ("ThermaCAM Researcher Professional 2.9") [51] la temperatura, la cual es dependiente del individuo y el aura (término dado por el oncológo a la luminosidad del contorno de la figura del paciente visualizada en la imagen térmica) se debe de calibrar pegada

al contorno de la figura, se selecciona la paleta de colores en Rain (paleta de colores RGB para una mejor visualización de la imagen térmica) para realizar el estudio basal(Imágenes térmicas tomadas antes del enfriamiento de las mamas que serán utilizadas como base para la comparación entre las imágenes de las mamas enfriadas con alcohol) y se cambia la paleta a Grey (paleta de color en escala de grises) para el estudio fisiológico (imágenes térmicas de las mamas enfriadas con alcohol), el paciente coloca sus brazos atrás de la nuca, sacando el pecho despegado de la silla, para obtener una mejor imagen, sólo se toma del cuello hacia abajo hasta antes de la cintura.

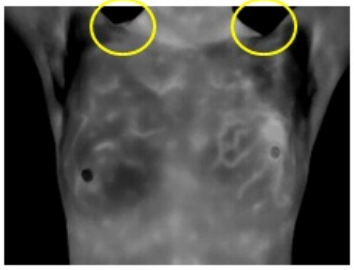

Figura 4.1: Ejemplo de triángulo formado en la imagen entre los hombros y el cuello, cuando los brazos del paciente son puestos por encima de la cabeza.

Se obtienen 6 imágenes por paciente, se puede observar un ejemplo de éstas en la figura 4.2. Las imágenes 1-3 conforman el estudio basal. La imágenes 4-6 el estudio funcional, en éste estudio las mamas son enfriadas con alcohol y se dejan reposar 2 minutos. A continuación se describe a las imágenes.

1. Se obtiene la imagen de las mamas de frente para la serie basal.
2. Para la obtención de la imagen, el paciente gira las piernas hacia su lado derecho y el tórax es girado $\frac{3}{4}$ hacia la derecha, capturando la imagen de la mama derecha para la serie basal.
3. Para la obtención de la imagen, el paciente gira las piernas hacia el lado izquierdo y el tórax es rotado $\frac{3}{4}$ a la izquierda capturando la imagen de la mama izquierda para la serie basal.
4. Se obtiene la imagen para la serie funcional de las mamas de frente.
5. Para la obtención de la imagen de la mama derecha de la serie funcional, el paciente gira las piernas hacia su lado derecho y el tórax es girado $\frac{3}{4}$ hacia la derecha.
6. Para la obtención de la imagen de la mama izquierda para la serie funcional, el paciente gira las piernas hacia el lado izquierdo y el tórax es rotado $\frac{3}{4}$ a la izquierda.

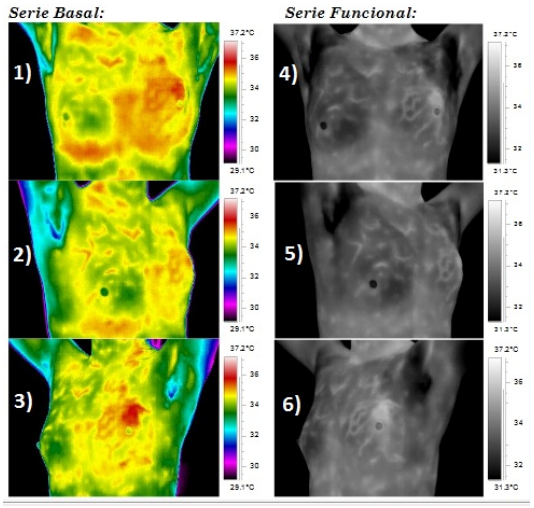


Figura 4.2: Ejemplo de serie basal y serie funcional, incluye la numeración de las imágenes como la realiza el oncólogo.

El oncólogo experto recomienda hacer la exploración mamaria después de la obtención de las imágenes térmicas, para no afectar el cambio de la temperatura que es registrado por la termografía, debido a que el contacto de las manos con las mamas puede hacer que aumente la temperatura de las mamas.

## 4.2. Descripción de formación de score termográfico

Existen 2 mediciones que son repetitivas porque a partir de ellas se obtienen algunos valores de las variables del score, estas mediciones son descritas a continuación.

1. **Cima térmica** es el punto con mayor temperatura en las mamas. Para realizar este cálculo se ocupan las imágenes de la serie funcional (ver figura 4.2), se selecciona de la paleta de color grey, también se selecciona de la barra de herramientas la isoterma (ver figura 4.3), se coloca en la escala de colores y se mueve hasta que se visualice el punto más caliente en color verde fluorescente, se baja la isoterma hasta que se observen en la imagen puntos ahuecados delineados de verdes fluorescente (esto indica los puntos donde existe una mayor temperatura), después es seleccionada otra isoterma que se visualizará en amarillo, se aplica encima de la isoterma verde y si lo ahuecado de la isoterma verde se pinta de amarillo, esto indica que es el centro caliente, probablemente sea la cima térmica, si se visualiza un solo centro caliente esa es la cima térmica,pero si se visualizan más de un centro caliente,

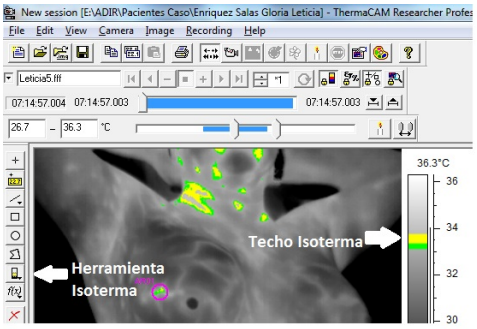


Figura 4.3: Ejemplo de herramienta Isoterma y techo Isoterma.

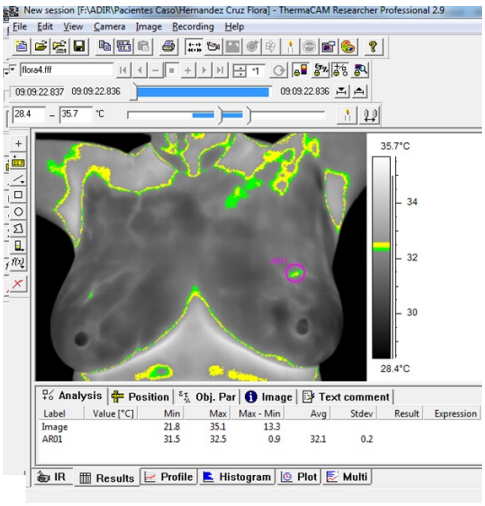


Figura 4.4: Ejemplo de cima térmica, seleccionada con la herramienta círculo.

se seleccionan con la herramienta de círculo hasta encerrar a todo lo verde fluorescente, se va a al apartado de resultados, el que tenga en "max.[el] valor más grande es la cima térmica, esto lo podemos observar en la figura 4.4.

2. **Hipertermia** se obtiene de las imágenes de la serie funcional (imagen 4,5 y 6), se establece la paleta de colores en Rain, y se mueve la escala de la temperatura hasta que se logren visualizar las venas que normalmente van a la cima térmica se puede ver en la figura 4.5.

Estos procedimiento son totalmente subjetivos, ya que depende de apreciación visual del analista de la imagen.

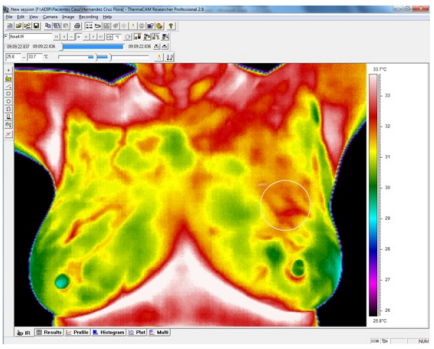

Figura 4.5: Ejemplo hipertermia, seleccionada con la herramienta círculo.

El score termográfico es formado por 16 variables, su obtención es descrita brevemente a continuación.

### 4.2.1. Asimetría (Asm)

En la imagen 1 y 4 (ver figura 4.2) se selecciona la paleta de color Rain, se ajusta la temperatura probando hasta que se visualice la vascularidad del paciente en la imagen, se selecciona de la barra de herramientas el polígono para seleccionar el área de la mama a la cual se le va a medir la temperatura, se mide desde el esternón hasta la axila formando un polígono que abarca toda la mama desde el surco (la parte inferior de la mama). De esta forma también se selecciona la otra mama, para realizar la comparación de ambas mamas como se puede ver en la figura 4.6. En el software, en la pestaña de resultados se observa en el cuadro un apartado llamado avg (promedio de la temperatura del área seleccionada), se obtiene el valor absoluto de la resta entre los avg (promedios) de las 2 mamas, ésto para la imagen 1 y 4; después de obtener los 2 valores absolutos de la imagen 1 y 4, se restan y se obtiene el valor absoluto. El valor asignado en el score es propuesto por el oncólogo experto de la siguiente manera: Si la diferencia es menor a 1 °C, se le asigna el valor de 5; si la diferencia ésta en un intervalo entre 1 °C y 2°C, se le asigna el valor de 10; si la diferencia es mayor a 2 °C se le asigna el valor de 15. Los valores que puede tomar Asm fueron designados por el oncólogo.

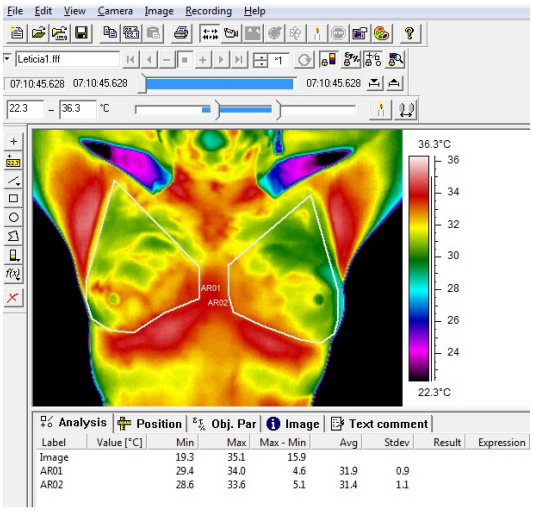


Figura 4.6: Ejemplo de la medición de las mamas, para obtener la diferencia de las temperaturas, se observa la selección de las mamas formando un polígono.

### 4.2.2. Red Termovascular (RT)

Para asignarle el valor a esta variable, se utilizan las imágenes 1,2 y 3 del estudio basal (ver figura 4.2), se selecciona la paleta de color en Rain, se selecciona de la barra de herramienta la Isoterma y se coloca en la escala de colores y se baja la escala hasta que se visualice la vascularidad en gris y se resalten en naranja o rojo las áreas con mayor temperatura. Si la vascularidad observada es abundante (dependiendo del número de venas que se visualicen) se le asigna el valor de 15, si es moderada el valor es 10 y si es leve el valor es 5, el valor es independiente para cada mama, en la figura 4.7 se muestra un ejemplo de red termovascular. Los valores que puede tomar RT fueron dados por el oncólogo.

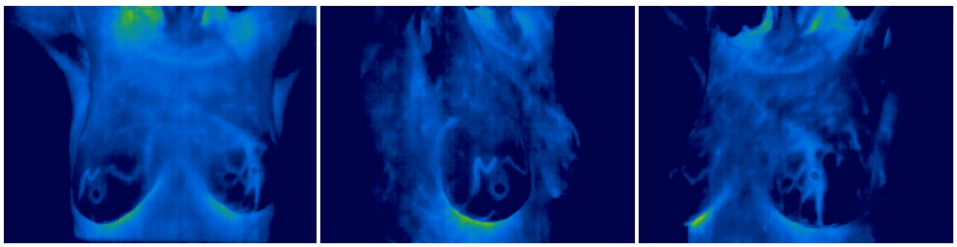

Figura 4.7: Ejemplo de Red termovascular, con imágenes 1,2 y 3. Para este ejemplo la red termovascular es predominante izquierda, con un valor de 15 para la mama izquierda, 5 para la mama derecha.

### 4.2.3. Patrón Curvilíneo (PCL)

Para asignarle el valor a esta variable se utilizan las imágenes 1,2 y 3 del estudio basal (ver figura 4.2), se selecciona la paleta de color Rain 900, se ajusta la escala de temperatura hasta que las imagen se tornen azul para resaltar la vascularidad, se ajusta la escala de temperatura hasta que se visualicen las axilas y el surco de color verde. Si lo que resalta es abundante se le da el valor de 15, si es moderado el valor es 10 y si es leve el valor es 5, el valor es independiente para cada mama, se puede ver un ejemplo en la figura 4.8. Los valores que puede tomar PCL fueron designados por el oncólogo.

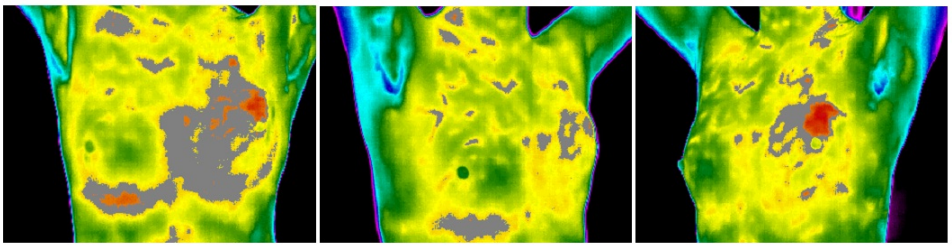

Figura 4.8: Ejemplo de patrón curvilíneo, con imágenes 1,2 y 3. Para este ejemplo el valor es de 15.

### 4.2.4. Porcentaje de Temperatura de mama (%)

A la imagen 1 (ver figura 4.2) se le aplica la paleta de color Rain, se ajusta la temperatura hasta que se visualice la vascularidad en la imagen, éste ajuste es dependiente de la apreciación del analista de las imágenes; se selecciona de la barra de herramientas el polígono, para seleccionar el área de la mama a medir, se mide desde el esternón hasta la axila formando un polígono que abarca toda la mama desde el surco, se selecciona la pestaña de histograma, en donde se muestra una gráfica de barras, en la parte inferior se debe seleccionar el área 1 (AR01) correspondiente a la mama derecha. El valor de esta variable es obtenido de la sumatoria de los porcentajes % (encontrado en el lado izquierdo de la ventana), se suman las 2 primeras barras de arriba hacia abajo. El resultado obtenido es redondeado, se debe de realizar lo mismo para la mama izquierda (AR02). Se muestra un ejemplo en la figura 4.9.

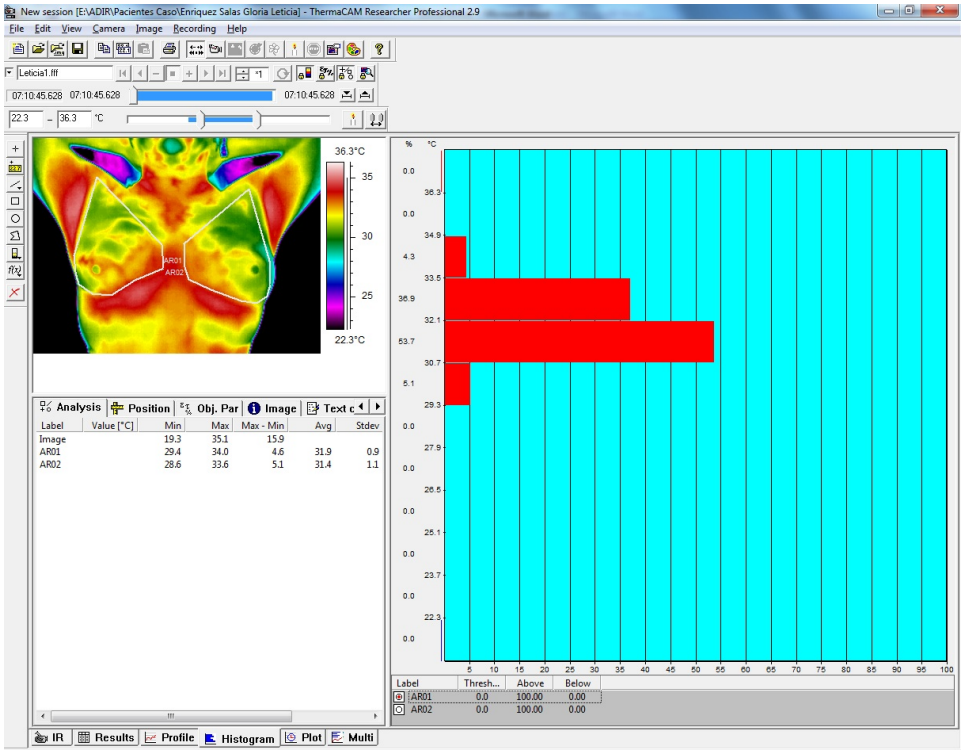


Figura 4.9: Ejemplo de la medición de las mamas y el histograma, para obtener el porcentaje de temperatura. El histograma mostrado es el de la mama derecha (AR01), en la cual 36.9 + 4.3 = 41.2 se redondea a 41.

### 4.2.5. Hipertermia funcional (HTF)

Una vez encontrada la cima térmica, si existe al menos una cima en la mama se le asigna el valor de 20 y si no existe cima el valor es 0, ver ejemplo en la figura 4.10. Los valores que puede tomar la variable fueron asignados por el oncólogo.

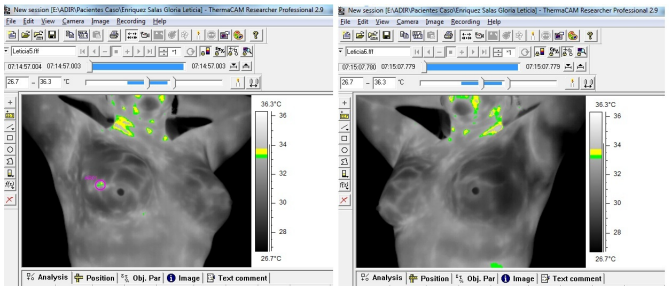

Figura 4.10: Ejemplo de hipertermia funcional. En la mama derecha se le asigna el valor de 20 (1 cima térmica), el valor de la mama izquierda es de 0 (no hay cima térmica).

### 4.2.6. Diferencia entre la cima de la mama y la otra mama con o sin cima (2c)

Ya que se obtuvo la cima térmica se selecciona de la barra de herramientas el çírculo", para medir la temperatura de la cima térmica, también se selecciona con la herramienta çírculo"la mama contralateral, a continuación se selecciona la pestaña de resultados, se obtiene el valor absoluto de la resta de los promedios (avg) de la cima y la contralateral. Si el valor obtenido es 1-10 se le asigna el valor de 10, si es de 11-15 es igual a 15, si es de 16-20 es de 20, si es mayor a 20 se le asigna 25. Si las 2 mamas tienen el mismo valor en la cima térmica al restarse es 0 y este es el valor que se le asignaría a esta variable.

### 4.2.7. Única hipertermia (única F)

Primero se obtiene la(s) hipertermia(s), si es una hipertermia se le asigna el valor de 40, en la mama en la que esté la hipertermia, si son 2 se le da el valor de 20, si son 3 se le asigna el valor de 10, si son más de 3 el valor asignado va a ser 5. Los valores de esta variable fueron designados por el oncológo.

### 4.2.8. Unilateral (1C)

Después de obtener la hipertermia se verifica si fue en una sola mama dándole el valor a esta variable de 40 para la mama con la única hipertermia, si fueron en las dos mamas a cada una se le asigna el valor de 20 que sumados nos dan 40. Los valores de 1C fueron designados por el oncólogo.

### 4.2.9. GAP

El GAP puede ser calculado de dos formas: la primera cuando existe un área de interés localizada por algún examen exploratorio médico o mastografía o ultrasonido, la segunda es cuando no existe un área de interés.
Para calcular el GAP con el área de interés se realiza de la siguiente manera: el GAP es la diferencia de las asimetrías del área de interés, para poderlo medir se ocupa alguna (s) las 6 imágenes (ver figura 4.2) dependiendo de en dónde esté ubicada el área de interés, se utiliza la herramienta polígono y se selecciona toda el área que ocupa la mama se deja fuera el pezón (por ser considerado unas de las áreas más calientes y podría afectar el resultado), y se hace una comparación entre las imágenes basales contra las fisiológicas. De la asimetría que hay entre ellas, si es mayor el valor de las imágenes fisiológicas que el de las basales esta diferencia es positiva y si es mayor la de las imágenes basales es negativa, cuando son iguales las diferencias el GAP es 0, ver ejemplo en figura 4.11. Para el segundo caso, cuando no existe un área de interés se toma como referencia la cima térmica y se obtiene la diferencia

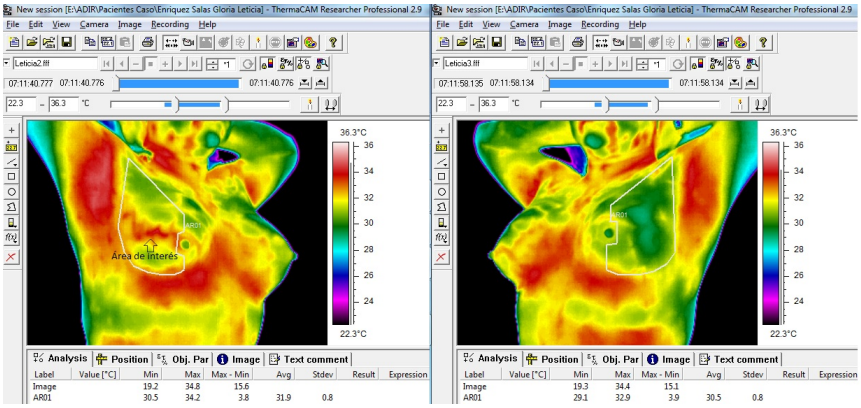


Figura 4.11: En la imagen se observa, como se mide la mama derecha e izquierda sin tomar en cuenta el pezón, la temperaturas promedios son mama derecha 31.9 mama izquierda 30.5, se restan y se tendría un valor de 1.4.

de temperaturas con respecto al tejido circundante (tejido que está alrededor de la posible lesión) y ese es el valor del GAP, ver ejemplo en figura 4.12.

### 4.2.10. Surco

A la imagen 1 (ver figura 4.2) se le aplica la paleta de color en Rain, se ajusta la temperatura hasta que se visualice la vascularidad del paciente en la imagen, se selecciona la isoterma que será visualizada en color gris, es colocada en la barra de temperaturas en la parte superior y se puede desplazar hasta visualizar el contorno inferior de las mamas en blanco o en gris, la que se visualice más es a la mama que se le asigna el valor positivo + y a la otra el valor negativo -, podemos ver un ejemplo en la figura 4.13. Los valores de esta variable fueron designados por el oncólogo.

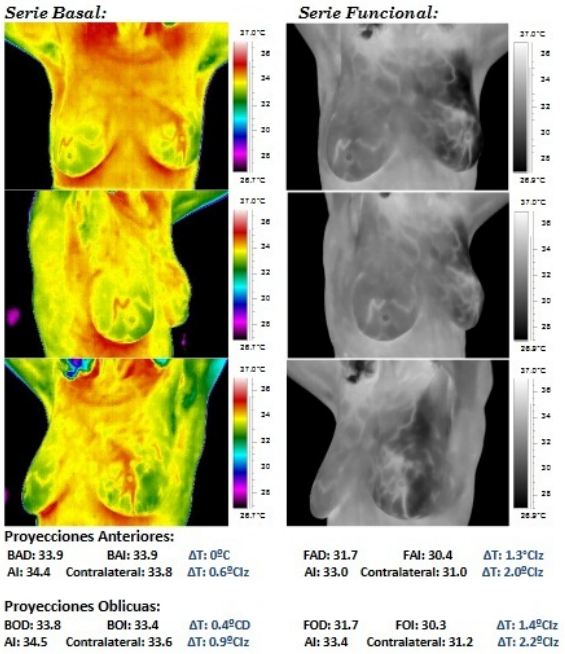


Figura 4.12: Así presenta el reporte el oncólogo, en él se muestran las 6 imágenes, 3 funcionales y 3 basales, para este ejemplo el área de interés es la mama izquierda, los valores son de las diferencias de la mama izquierda de 1.3 y 1.4, siendo el más alto el 1.4 este es el valor del GAP.

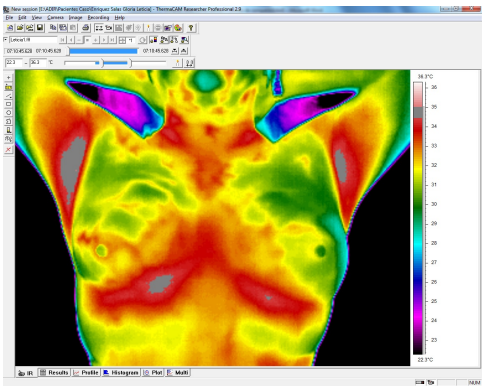

Figura 4.13: En la imagen se observa el surco de las mamas, el más marcado es el de el lado derecho, el valor para la mama derecha es -", para la mama izquierda ".

### 4.2.11. Pin point (PP)

Una vez encontrada la hipertermia, se observa si tiene algún vaso vascular que señale la hipertermia o vaya directamente a ella, en caso de haberlo es positivo -"por el contrario negativo ", ver figura 4.14. Los valores de esta variable fueron designados por el oncólogo.

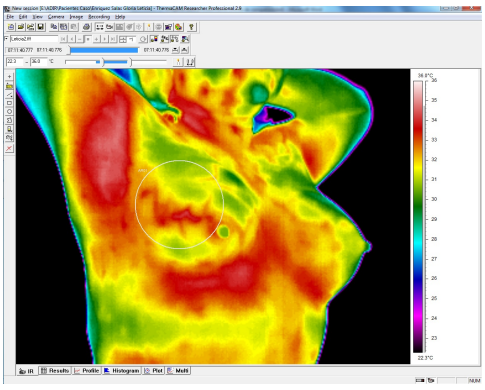

Figura 4.14: Ejemplo de Pin point en mama derecha, para este ejemplo es positivo, esta encerrado en un círculo para una mejor apreciación.

### 4.2.12. Centro Caliente (CC)

Después de haber localizado la cima térmica si el isoterma verde se pinta en la parte ahuecada de amarillo, quiere decir que el centro si es caliente y se le asigna el valor de -", en caso de que ésto no pase el valor es negativo ", ver figura 4.15. Los valores que toma esta variable fueron designados por el oncólogo.

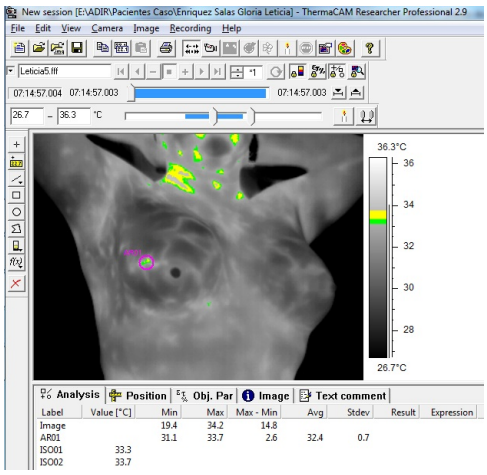


Figura 4.15: Ejemplo de Centro Caliente, es seleccionada la cima térmica, mostrándose como se pinta de amarillo su interior.

### 4.2.13. Forma Irregular de la Cima (FD)

Una vez encontrada la cima térmica se observa su forma, en caso de ser irregular (no apariencia geométrica). El valor de FD es positivo -", si es de forma regular es negativo ", ver figura 4.16. El valor de esta variable depende totalmente de la apreciación de técnico que analiza la imagen. Los valores que toma FD fueron designados por el oncólogo.

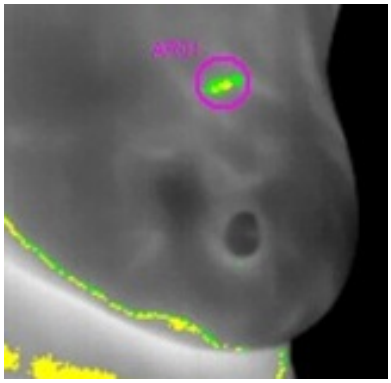

Figura 4.16: Ejemplo de forma irregular en la cima, para este caso el técnico le asigno el valor de -.ª la variable.

### 4.2.14. Histograma en forma de triángulo (H)

Se selecciona una de las imágenes fisiológicas 4,5 ó 6 (ver figura 4.2), dependiendo de en cual se haya encontrado la cima térmica, se selecciona la paleta de color a Rain, se aplica la isoterma, se baja el cuadro de la

isoterma pintándose en rojo el contorno de la cima térmica y el interior de gris, se selecciona la cima térmica con la herramienta çírculoçubriendo en su totalidad el área gris, se selecciona la pestaña de histograma, observándose una gráfica de barras, si las barras en la gráfica se visualiza en forma de un triángulo, el valor de esta variable será positivo -", si no se visualiza el valor es negativo ". Esta variable subjetiva, ya que depende de la apreciación de la persona que analiza la imagen, ver ejemplo en la figura 4.17. Los valores que se le pueden asignar a H fueron dados por el oncólogo.

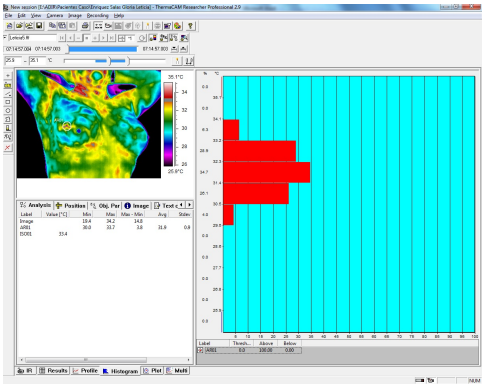

Figura 4.17: Ejemplo de histograma en forma de triángulo, es seleccionada la cima térmica, mostrándose el histograma que asemeja una forma de triángulo, el valor es positivo -".

### 4.2.15. Axila (Ax)

Se utilizan las imágenes 2 y 3 (ver figura 4.2), se selecciona de la barra de herramientas la herramienta polígono, se forma un rombo abarcando toda la axila, se selecciona el apartado de resultados, se comparan los promedios de las 2 axilas y el promedio más alto, es la axila positiva + y la otra axila es negativa-, en caso de que los promedios sean iguales las dos axilas son negativas, ver figura 4.18. Los valores que puede tomar la variable Ax fueron designados por el oncólogo.

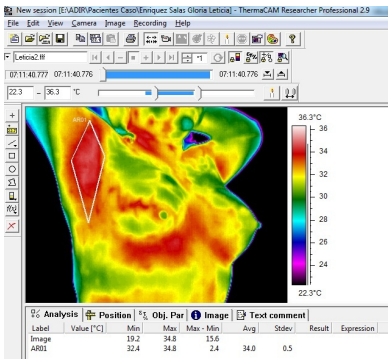

Figura 4.18: Ejemplo de medición de Axila seleccionada en forma de rombo.

### 4.2.16. Perfil Alterado (PF)

Se tienen que observar la imágenes 1,2 y 3 (ver figura 4.2), en caso de encontrar una alteración en el surco o en el perfil izquierdo o en el perfil derecho, se anotará su grado de alteración. Ésta medida es subjetiva y depende de la apreciación del analista de la imagen. La variable PF puede tomar 4 valores: negativo (-) cundo no se aprecia ninguna alteración en las mamas, se le da el valor de (+) cuando la alteración es leve, cuando se observa una alteración moderada se le asigna (++), si la alteración es muy evidente y se aprecia severa se le asigna el valor (+++). Los valores que toma esta variable dependen de la apreciación de la persona que analiza la imagen, se puede ver un ejemplo en la figura 4.19. Los valores que PF puede tomar fueron dados por el oncólogo.

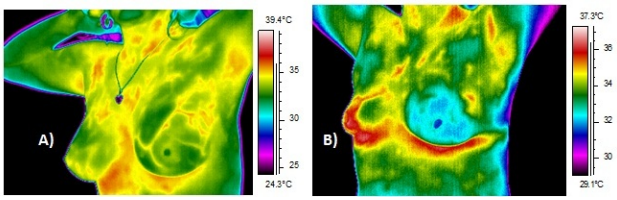


Figura 4.19: Ejemplo de perfil alterado, la imagen A) tiene un perfil (-), la imagen B)tiene un perfil alterado (++).

### 4.2.17. Fórmula para obtención del score

Para obtener el valor del score se realiza una sumatoria de las siguientes variables termográficas: Asimetría (ASM), Red Termovascular (RT), Patrón Curvilíneo (PCL), Porcentaje de Temperatura de mama %, Hipertermia funcional (HTF), Diferencia entre la cima de la mama y la otra mama con o sin cima (2c), Única hipertermia (única F),Unilateral (1C). La fórmula es la

siguiente.

$$Score = \sum_{i=1}^{n} x_i \tag{4.1}$$

donde $x_i$ son los valores de las variables termográficas que van desde $i = 1, 2, \ldots, n$ y $n$ es el número de variables termográficas, que para nuestro caso $n = 8$, porque se utilizan los valores de las primeras 8 variables termográficas.

## 4.3. Glosario de términos

Hay algunos términos que el oncólogo utiliza para describir las acciones que realiza, algunos de ellos son los siguientes:

1. Aura. Es la luminosidad del contorno de la figura del paciente visualizada en la imagen, esta luminosidad se debe de calibrar, para que se observe pegada al contorno de la figura del paciente, ver figura 4.20. La calibración de la imagen es subjetiva porque depende de la apreciación del analista de la imagen.

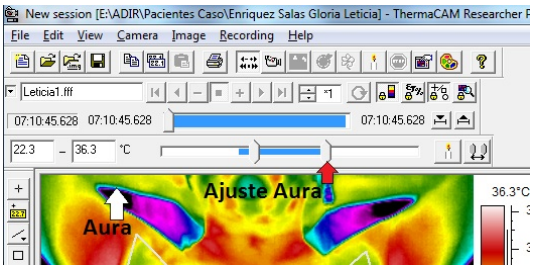


Figura 4.20: Ejemplo de Aura y en dónde se realiza el ajuste.

2. SPAM. Es el ajuste que se realiza a la temperatura corporal del paciente hasta que se visualiza la vascularidad, la temperatura depende de cada persona y ésta es calibrada por la apreciación del analista de la imagen por lo tanto es de manera subjetiva, ver figura 4.21.

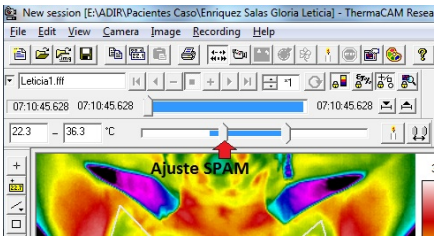


Figura 4.21: Ejemplo de Ajuste de SPAM.

3. Techo de Isoterma. Se utiliza para poder obtener el centro caliente y la cima térmica, este nombre fue dado por el oncólogo. Primero se selecciona la herramienta isoterma, después aparece a un costado una barra de desplazamiento para poder calibrar la isoterma, después se vuelve a seleccionar la herramienta isoterma y ésta aparece por encima de la otra isoterma. A ésto el oncólogo le llama techo, al mover las isotermas por la barra de desplazamiento se pueden pintar el centro de los puntos calientes, se puede ver en la figura 4.3.

# Capítulo 5

# Materiales y Métodos

## 5.1. Descripción del conjunto de datos

El conjunto de datos fue recopilado por el oncólogo de 100 pacientes con sospecha de cáncer por mastografía y/o ultrasonido, se obtuvieron un total de 98 casos: 77 (78.57 %) de los casos con pacientes con cáncer de mama y 21 (21.43 %) de casos con pacientes sanos. Todos los resultados enfermo/sano fueron confirmados con biopsia, que es considerado el método de oro en la detección de cáncer de mama. Derivado del análisis de las imágenes térmicas se obtienen 16 variables, de las cuales 7 variables son nominales, 2 variables numéricas y 7 dicotómicas (+,-) o binarias (0, 1). Las variables que se suman para darle el valor al score son: asimetría,red termovascular, patrón curvílineo, porcentaje de temperatura, hipertermia funcional, 2c y 1c. Las 8 variables restantes en conjunto con el score ayudan al oncólogo a poder dar un pre-diagnóstico. En la Tabla 5.1 se dan detalles del nombre y el tipo de cada una de estas variables.

Tabla 5.1: Nombre, descripción y tipos de variables del estudio termográfico mamario.

| **No.** | **Variable** | **Definición** | **Tipo de variable** |
|---|---|---|---|
| 1 | Asimetría | Grados de diferencia (en Celsius) entre mama derecha e izquierda | Nominal (rango [1-3]) |
| 2 | Red termovascular | La cantidad de venas con temperatura más alta | Nominal (rango [1-3]) |
| 3 | Patrón curvilíneo | El área más caliente detro de la mama | Nominal (rango [1-3]) |
| 4 | Porcentaje de temperatura | Porcentaje de calor de la mama con mayor temperatura | Numérico |

| No. | Variable | Definición | Valor de variable |
|---|---|---|---|
| 5 | Hipertermia funcional | Punto más caliente de la mama | Nominal [0 ó 40] |
| 6 | 2c | Diferencia de grados entre los puntos más calientes de la mama | Nominal (rango [1-4]) |
| 7 | Única F | Cantidad de puntos más calientes | Nominal (rango [1-4]) |
| 8 | 1c | Punto más caliente en una sola mama | Nominal [20 ó 40] |
| 9 | GAP | Es la diferencia de los promedios de temperatura entre mama derecha e izquierda | Numérico |
| 10 | Surco | Surco debajo de las mamas | Binario |
| 11 | Pinpoint | Las venas que van a los puntos más calientes de las mamas | Binario |
| 12 | Centro caliente | El centro de la zona más caliente | Binario |
| 13 | Forma irregular | Geometría del centro caliente | Binario |
| 14 | Histograma | Histograma en forma de un triángulo isósceles | Binario |
| 15 | Axila | Diferencia de grados entre las 2 axilas | Binario |
| 16 | Perfil Alterado | Visualmente el perfil alterado de la mama | Binario |
| 17 | Score | Si la suma de las primeras 8 variables $\leq$ 160 (1) y si es $<$ 160 (0) | Binario |
| 18 | Clase | Cancer/no cancer | Binario |

Son utilizadas otras variables para análisis estadístico que no forman parte del estudio termográfico mamario éstas son. Las variables de patología, a partir de ellas se sabe si el paciente tiene cáncer o no, se da una breve descripción en la tabla 5.2.

Tabla 5.2: Nombres, descripción y tipos de variables de patología

| No. | Variable | Definición | Tipo de variable |
|---|---|---|---|
| 1 | TamañoD | Tamaño del tumor discretizado, 1 $<$ a 2 cm, 2 de 2cm - 5cm y 3 $>$ 5cm | Nominal (rango [1-3]) |

| No. | Variable | Definición | Tipo de variable |
|---|---|---|---|
| 2 | RHP | Tipos de lesiones cancerosas y no cancerosas | Nominal (rango [1-7] cáncer, [8-16] no-cáncer) |
| 3 | SBR grado | Grado de malignidad entre mayor seá el grado el cáncer es más agresivo, grado = 0 sin cáncer | Nominal (rango [0-3]) |

## 5.2. Estadística Descriptiva

La Estadística es la disciplina que se ocupa de i) la recolección, organización, resumen y análisis de datos, y ii) la obtención de inferencias a partir de un volumen de datos cuando se examina sólo una parte de éstos [52, 53]. Otra forma de definir la estadística es a través de las medidas descriptivas calculadas a partir de los datos de una muestra.

Las variables son una característica de la muestra o de la población que se analiza en un estudio estadístico, éstas pueden ser cualitativas o cuantitativas.

*Variable cualitativa.* Es aquella variable que no puede ser medida de forma usual, que se puede expresar normalmente por una palabra y no por números, por ejemplo estado civil, nacionalidad, sexo.

*Variable cuantitativa.* Es aquella que se expresa numéricamente, por ejemplo edad, peso. Las variables cuantitativas pueden ser discretas o continuas. Las variables discretas son aquellas que sólo pueden tomar determinados valores, número enteros, por ejemplo (el número de hijos por familia, número de empleados de una empresa). Las variables continuas son aquellas que toman cualquier valor dentro de un intervalo dado (por ejemplo la estatura).

En la estadística descriptiva hay medidas descriptivas las cuales se pueden calcular a partir del conjunto de datos de muestra, entre estas medidas encontramos a las medidas de tendencia central y las medidas de dispersión[53].

### 5.2.1. Medidas de tendencia central

Dentro de las medidas de tendencia central se encuentra la media aritmética, la mediana y la moda, las cuales serán descritas a continuación.

***Media aritmética***. Es la medida de tendencia central más conocida y comúnmente llamada "promedio". La media se calcula sumando todos los valores de una población o muestra y dividiendo entre el número de valores

sumados. La fórmula de la media es la siguiente.

$$\mu = \frac{\sum_{i=1}^{n} x_i}{n} \tag{5.1}$$

donde $x_i$ es un número específico de valores de una muestra que van desde $i = 1, 2, \ldots, n$ y $n$ es el número de elementos.

***Mediana***. La mediana de un conjunto de valores es aquel valor que divide al conjunto en dos partes iguales, de forma que el número de valores mayores o iguales a la mediana es igual al número de valores menores o iguales a ésta. Si el número de elementos de la muestra es impar, la mediana es el valor medio o central siempre y cuando todos estos elementos sean ordenados por magnitud, cuando el número de elementos es par, no existe un valor medio único, si no que existen 2 valores medios. Para este caso, la mediana corresponde a la media de esos dos valores centrales, cuando todos los valores son ordenados. Es decir, la mediana del conjunto de datos es la $(n+1)/2-$ésima observación, cuando las observaciones han sido ordenadas [54].

***Moda***. La moda de un conjunto o muestra es aquel valor que ocurre con mayor frecuencia. Si todos los valores son diferentes, no hay moda. Un conjunto de valores puede tener más de una moda [52].

En el capítulo de resultados, son calculadas para nuestro conjunto de datos estas medidas, a partir de éstas medidas podemos observar el comportamiento del conjunto de datos.

### 5.2.2. Medidas de dispersión

La dispersión de un conjunto de valores se refiere a la variedad que muestran éstos. Una medida de dispersión nos indica la variabilidad que está presente en el conjunto de datos. Si todos los valores son iguales no hay dispersión, pero si los datos no son iguales entonces existe dispersión. Se dice que una dispersión es pequeña cuando a pesar de que los valores son diferentes, pero son cercanos entre sí [55]. A continuación se definen algunas medidas de dispersión.

***Varianza***. Cuando los valores de un conjunto de observaciones se encuentran ubicados cerca de la media, la dispersión es menor que cuando están esparcidos. Por lo tanto es posible medir la dispersión en función de en cuanto están esparcidos los datos con respecto a su media. Esta medida se efectúa mediante la varianza. Para calcular la varianza de una muestra de valores, se resta la media a cada uno de los valores individuales, las diferencias se elevan al cuadrado y después se suman entre sí. Esta suma de desviaciones

elevadas al cuadrado de los valores con respecto a la media se divide entre el tamaño de la muestra, menos 1, para obtener la varianza de la muestra [53].

$$S^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} \tag{5.2}$$

donde $xi$ es el número de valores que van desde $i = 1, 2, \ldots, n$, $n$ es el número de elementos y $\bar{x}$ es la media.

***Desviación estándar***. La varianza descrita previamente representa unidades al cuadrado, por lo que no es una medida expresada en términos de las unidades originales. Para obtener la medida de dispersión en unidades originales, se obtiene la raíz cuadrada de la varianza [53]. La desviación estándar de una muestra se obtiene con la siguiente fórmula.

$$\sigma^2 = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}} \tag{5.3}$$

## 5.3. Correlaciones

En ocasiones se requiere saber si existe algún tipo de relación entre 2 variables aleatorias. Una forma de determinar esta relación podría ser dibujando puntos en un plano de coordenadas $x$ y $y$, así se obtendría un conjunto de puntos los cuales nos indicarían visualmente si existe algún tipo de relación (por ejemplo lineal, parabólica, exponencial, etc).

Para poder cuantificar la intensidad de la relación lineal que existe entre dos variables. El parámetro que nos da tal cuantificación es el coeficiente de correlación lineal de Pearson $r$, cuyo valor oscila entre -1 y +1, éste se puede calcular mediante la siguiente fórmula [53, 52].

$$-1 \leq r = \frac{cov(x,y)}{S_x S_y} = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) * (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} * \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}} \leq +1 \tag{5.4}$$

Donde $cov(x, y)$ es la variación conjunta entre dos variables aleatorias a esto se le llama covarianza, $x_i, y_i$ son los valores de las variables $x, y$ que van desde $i = 1, 2, \ldots, n$ y $n$ es el número de elementos de la muestra, $\bar{x}, \bar{y}$ son las medias de las variables $x, y$. Como ya se mencionó previamente, $r$ puede tomar valores que van de -1 a +1, cuando $r = -1$ existe una relación entre las 2 variables lineal perfecta y negativa (significa que los valores bajos de

una de las variables corresponden a los valores altos de la otra variable), si $r = 1$ la relación es perfectamente lineal y positiva (los valores bajos de una de las variables corresponde a los valores bajos de la otra variable), cuando $r = 0$ no existe relación lineal (los valores de las variables no tienen un orden), como se puede observar en las figuras 5.1- 5.4.

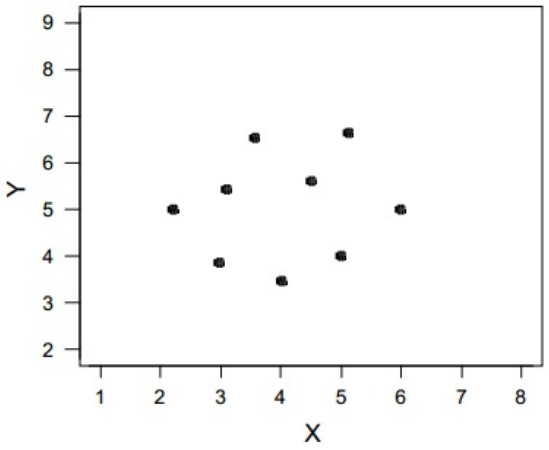


Figura 5.1: Variables no correlacionadas $r = 0$

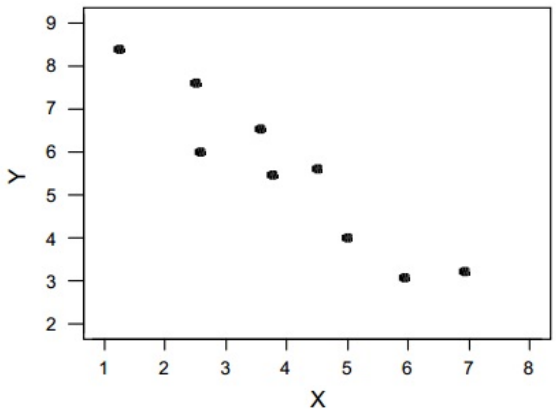


Figura 5.2: Correlación lineal negativa $r = -1$

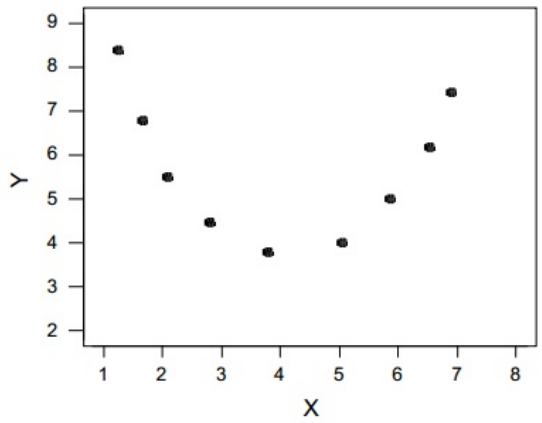


Figura 5.3: Correlación no lineal $r = 0$

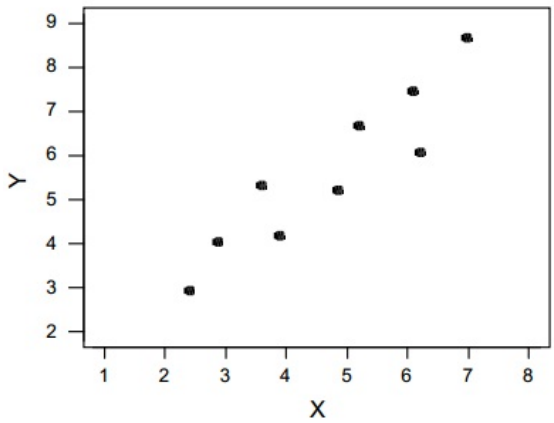


Figura 5.4: Correlación lineal positiva $r = 1$

## 5.4. Clasificadores

Existen objetos que son agrupados en clases, donde se tiene de cada clase una muestra de objetos llamadas instancias o ejemplos; que se sabe que pertenecen a esta clase (Clasificación supervisada) [25, 56], pero surge un problema, el poder establecer a que clase pertenece un nuevo objeto.

Los algoritmos de clasificación tienen como objetivo determinar a qué clase pertenece una instancia, que es descrita por un conjunto de atributos a una o varias clases, basándose en la información proporcionada por un conjunto de datos previamente clasificados (conjunto de entrenamiento)[10]. Para este trabajo contamos con un conjunto de datos descrito previamente en la sección 5.1. Al conjunto de datos se le requiere clasificar. Para clasificar los datos, se utilizaron los algoritmos descritos a continuación. Se desea en-

contrar el algoritmo que ofrezca un mayor número de instancias clasificadas correctamente.

### 5.4.1. K-NN

#### 5.4.1.1. Introducción K-NN

El algoritmo K-NN clasifica casos basándose en su parecido con respecto a otros casos. Los casos parecidos están próximos y los que no lo son están alejados entre sí, lo que quiere decir que la distancia entre los dos casos es una medida de diferencia. Los casos que están próximos entre sí se denominan vecinos. Cuando se presenta un nuevo caso, se calcula su distancia con respecto a los casos del modelo y las clasificaciones donde los casos sean los más parecidos (los vecinos más próximos) se cuadran y el nuevo caso es incluido en la clase que contiene el mayor número de vecinos más próximos. El número de vecinos que se van a examinar esta dado por el parámetro $k$.

**Distancia Euclideana**

El criterio de comparación principalmente usado en los métodos de vecindad (K-NN) es la distancia, existen diferentes formas de medirlo entre ellas encontramos: distancia de Manhattan, Chebychev, Mahalanobis, distancia del coseno, distancia la función delta. La distancia más utilizada es la Euclideana, expresada de la siguiente manera.

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{5.5}$$

donde $xi$ y $yi$ son puntos que van desde $i = 1, 2, \ldots, n$ y $n$ es el número de dimensiones.

K-NN se detalla en el algoritmo 1.

**Algoritmo 1** K-NN

**Require:** $D$= $\{(x_1, c_1), \ldots, (x_n, c_n)\}$ {instancias ya clasificadas}, $x = (x_1, \ldots, x_n)$ {nuevo caso a clasificar}
**for** todo objeto ya clasificado $(x_i, c_i)$ **do**
  Calcular $di = d(x_i, x)$ {distancia Euclideana}
**end for**
Ordenar $d_i (i = 1, \ldots, n)$ {en orden ascendente}
Quedarnos con los $k$ casos $D_x^k$ ya clasificados más cercanos a $x$
Asignar a $x$ la clase más frecuente en $D_x^k$

#### 5.4.1.2. Ventajas y Desventajas K-NN

Las ventajas y desventajas son las encontradas en la literatura especializada [57, 58, 59].

**Ventajas**

- Aprende conceptos complejos a través de una función sencilla (distancia entre los instancias).
- Se puede extender la función, por ejemplo, mediante regresión para atributos continuos.
- Es muy tolerante al ruido.

**Desventajas**

- El costo computacional de encontrar $k$ mejores vecinos es alto.
- Se tiene que encontrar el valor óptimo para k (es dependiente del conjunto de datos).
- El rendimiento del algoritmo baja si el número de ejemplos aumenta.
- Carece de interpretabilidad, porque solo nos indica a que clase pertenece una instancia.

### 5.4.2. ID3

#### 5.4.2.1. Introducción ID3

ID3 es un algoritmo de aprendizaje supervisado desarrollado por J. Ross Quinlan en 1986 [60]. ID3 construye un árbol de decisiones mediante el método top-down (de lo más general a lo más específico). Dado un conjunto de datos de entrenamiento trata de probar cada atributo en un nodo del árbol. El algoritmo elige el mejor atributo utilizando una heurística llamada ganancia de información, así coloca dicho atributo en el nodo. La ganancia de información mide qué tan bien un determinado atributo separa los ejemplos de entrenamiento de acuerdo a su objetivo de clasificación.

**Entropía**

Es una medida de la teoría de la información, la entropía se define como una medida de incertidumbre promedio, la cual se calcula a partir de la probabilidad de ocurrencia de cada uno de los eventos [61], que caracteriza el ruido de una colección arbitraria de ejemplos [62]. Si el atributo objetivo

toma valores diferentes para $c$, entonces la entropía de $S$ es relativa para cada valor de $c$ es definida como:

$$Entropa(S) = \sum_{i=1}^{c} -p_i \, log_2 \, p_i \tag{5.6}$$

Donde $p_i$ es la proporción/ la probabilidad de que $S$ pertenezca a la clase $i$. El logaritmo en base 2 es porque la entropía es una medida de la longitud de codificación y tal medida es esperada en bits.

**Ganancia de información**

La ganancia de información se define como la reducción de la entropía causada por particionar un conjunto de ejemplos S con respecto a un atributo A; expresado en la siguiente fórmula.

$$Ganancia(S, A) = Entropa(S) - \sum_{v \in Valores(A)} \frac{S_v}{S} Entropa(S_v) \tag{5.7}$$

donde los $Valores(A)$ es el conjunto de todos los posibles valores del atributo $A$ y $S_v$ es un subconjunto de $S$ para los cuales el atributo $A$ tiene el valor $v$. Esta medida se puede utilizar para clasificar los atributos y construir el árbol de decisión, en donde el nodo raíz es el atributo con mayor ganancia de información.

#### 5.4.2.2. Algoritmo ID3

La decisión central de ID3 versa sobre qué atributo colocará en cada nodo del árbol. Para poder tomar la decisión se deben de cuantificar las bondades de un atributo; para realizarlo podemos utilizar la cantidad de información que éste proveerá, como lo define la teoría de la información por Claude E. Shannon [63].
Teoría de la información es la cantidad media de información (en bits) necesaria para codificar la clasificación de un ejemplo [63]. A continuación se presenta en el Algoritmo 2 a ID3.

#### 5.4.2.3. Ventajas y Desventajas ID3

Las ventajas y desventajas de este algortimo [64, 65].

**Ventajas**

- Es sobre los árboles de decisión que más investigación se ha realizado.
- Facilita la interpretación de la decisión adoptada.
- Proporciona una buena comprensión del conocimiento utilizado en la toma de decisiones.

**Desventajas**

- Con atributos continuos, no clasifica correctamente los ejemplo dados.
- La selección de las variables que ocuparán el nodo, se ve sesgada por el número de valores diferentes que posee la variable.
- Para problemas donde hay muchas variables y éstas a su vez toman varios valores, se generan arboles grandes (el conjunto de reglas es muy alto).

**Algoritmo 2** ID3

**Require:** Ejemplos, Atributos, Clases
**if** Ejemplos $= \emptyset$ **then**
  **return** Árbol raíz formado por nodo clase $= Valor_{Predefinido}$
**end if**
**if** $\forall$ Ejemplos $\in C_i$ **then**
  **return** Árbol raíz formado por nodo clase $= C_i$
**end if**
**if** Atributos $= \emptyset$ **then**
  **return** Árbol raíz formado por nodo clase = Valor-Mayoría(Ejemplos)
**else**
  $Atributo_{Mejor}$ = Escoger-Atributo(Atributos, Ejemplos)
  Árbol = Nuevo árbol raíz formado por nodo $Atributo_{Mejor}$
  **for** $i = 1$ $n$ valores de $Atributo_{Mejor}$ **do**
    $Ejemplo_i = e \in Ejemplost.q.Valor(e) = v_i$
    $Valor_{Predefinido}$ =Valor-Mayoría(Ejemplos)
    Subárbol=Aprendizaje-AD($Ejemplos_i$, $Atributos - \{Atributo_{Mejor}\}$, $Valor_{Predefinido}$)
    Añadir una rama a Árbol con etiqueta $= v_i$
    subárbol = Subárbol
  **end for**
**end if**
**return** árbol

### 5.4.3. C4.5

#### 5.4.3.1. Introducción C4.5

El clasificador $C4{,}5$ es una extensión del algoritmo ID3, para resolver algunas deficiencias de éste, como el favorecer a los atributos con un mayor número de valores diferentes [60], fue propuesto por Quinlan en 1993 [66]. Este algoritmo genera un árbol de decisión a partir de los datos a través de particiones realizadas recursivamente, el árbol es construido mediante la estrategia de profundidad-primero (*depth-first*).

El algoritmo $C4{,}5$ utiliza una técnica heurística conocida como proporción de ganancia (*gain ratio*), la cual favorece aquellos atributos que, en igualdad de ganancia, separen los datos en menos clases. A continuación se definen algunas fórmulas para poder llegar a la proporción de ganancia.

$$ParticinInformacion(S, A) = -\sum_{i=1}^{C} \frac{|S_i|}{|S|} log_2 \left( \frac{|S_i|}{|S|} \right) \tag{5.8}$$

$$ProporcinGanancia(S, A) : \frac{Ganancia(S, A)}{ParticionInformacion(S, A)} \tag{5.9}$$

donde la fórmula 5.8 es la partición de la información (*split information*) la cual es el cálculo del promedio ponderado de la información utilizando la proporción del conjunto de ejemplos, donde $S_i$ es un subconjunto de $S$ y es el número de valores que puede tomar el atributo $A$ con respecto a la clase $C$. La fórmula 5.9 es la proporción de la ganancia (*gain ratio*); divide la ganancia de la información entre la partición de la información de un conjunto de ejemplos $S$ dado un atributo $A$. Esto penaliza a diferencia de ID3 la partición de información con muchos valores.

C4.5 se detalla en el Algoritmo 3.

#### 5.4.3.2. Ventajas y Desventajas C4.5

Las ventajas y desventajas son las encontradas en la literatura especializada [64, 65].

**Ventajas**

- Se puede evitar el sobre-ajuste de los datos (overfitting).
- Se puede determinar qué tan profundo debe crecer el árbol de decisión.
- Reduce errores en la poda o corte (prunning).
- Se puede condicionar la post-poda.

**Algoritmo 3** C4.5

**Require:** $R$: Conjunto de atributos no clasificadores, $C$: Atributo Clasificador, $S$: Conjunto de entrenamiento,

**Ensure:** Devuelve un árbol de decisión

1: **if** $S$ está vació **then**
2: **return** Único Nodo con valor Falla {Forma el nodo Raíz}
3: **end if**
4: **if** Todos los $S$ tienen el mismo valor para $C$ **then**
5: **return** Único Nodo valor {único nodo para todos}
6: **end if**
7: **if** $R$ está vació **then**
8: **return** Único Nodo con el valor más frecuente del atributo clasificador en los registros de $S$ {habrá errores, es decir, registros que no estarán bien clasificados}
9: **else**
10: $D \leftarrow$ atributo con mayor proporción de ganancia $(D, S)$ entre los atributos de $R$
11: Sean $\{d_j | j = 1, 2, \ldots, m\}$ los valores del atributo $D$
12: Sean $\{d_j | j = 1, 2, \ldots, m\}$ los subconjuntos de $S$ correspondientes a los valores de $d_j$ respectivamente
13: **return** un arbol con la raíz nombrada como $D$ y con los arcos nombrados $d_1, d_2, \ldots, m$ que van respectivamente a los árboles $C4{,}5(R - D, C, S_1), C4{,}5(R - D, C, S_2), C4{,}5(R - D, C, S_m)$
14: **end if**

- Maneja atributos con valores continuos y con valores perdidos.
- Las reglas de decisión son simples y legibles, por tanto la interpretación de los resultados es directa e intuitiva.
- Es robusto frente a datos atípicos u observaciones mal etiquetadas.
- Es computacionalmente rápido.

**Desventajas**

- Aprende muy bien los datos de entrenamiento hasta el grado de aprender el ruido de los mismos, esto se vuelve un problema cuando trata de generalizar con los datos de prueba, esto se puede evitar de dos formas : 1) frenando el crecimiento del árbol antes de que llegue a clasificar perfectamente los datos de entrenamiento, 2) dejando crecer el árbol y después realizar una poda o corte.

### 5.4.4. AdaBoost

#### 5.4.4.1. Introducción AdaBoost

AdaBoost (*Adapting boosting*) es una variante del método boosting , el cual cae dentro de los métodos ensambladores [67]. Un método ensamblador consiste en combinar un conjunto de métodos para obtener una mayor precisión, la cual estos clasificadores de forma individual no obtendrían.

El algoritmo AdaBoost fue desarrollado por Schapire y Freund en 1996 [67, 68]. AdaBoost tiene una función que consiste en penalizar a los ejemplos mal clasificados obligando al clasificador débil a enfocarse en estos ejemplos, los cuales le son más difíciles de clasificar. El objetivo es minimizar el error del clasificador, mediante la asignación de mayor peso a los ejemplos mal clasificados y menor peso a los ejemplos clasificados correctamente.

Adaboost se detalla en el Algoritmo 4.

#### 5.4.4.2. Ventajas y Desventajas AdaBoost

Las ventajas y desventajas son las encontradas en la literatura especializada [69, 70, 71].

**Ventajas**

- Solo se requiere un único parámetro ($T$) el número de iteraciones.
- No requiere ningún conocimiento previo sobre el clasificador débil y así puede combinar de forma flexible cualquier clasificador para encontrar hipótesis débiles.

**Algoritmo 4** AdaBoost

**Require:** $S = (x_i, y_i)\, i = 1, \ldots, n$; $H$ = base de entrenamiento donde $x_i \in X, y_i \in Y = -1, +1$ {muestras de entrenamiento}

**Ensure:** $H(x) = sign\left(\sum_{t=1}^{T} \alpha h_t(x)\right)$ {hipótesis final}

1: $w_i(i)\frac{1}{n}$
2: **for** $t = 1 \rightarrow T$ {donde $T$ = es el número de iteraciones} **do**
3: Obtención de clasificadores débiles $h_t : X \rightarrow -1, +1$ con error, $\varepsilon_t = Pr[h_t(x_i) \neq y_i]$
4: Se elige $\alpha_t = \frac{1}{2} ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$
5: $w(i)_{t+1} = \frac{w(i)_t}{z_t} \times \begin{cases} e^{-\alpha} & \text{si } h(x_i)_t = y_i \\ e^{\alpha} & \text{si } h(x_i)_t \neq y_i \end{cases}$ {actualizar pesos o distribuciones}
6: **end for**

- Viene con un conjunto de garantías teóricas dadas por los clasificadores que está ensamblando.

**Desventajas**

- Es incapaz de manejar clasificadores débiles con una tasa de error mayor a $\frac{1}{2}$.

### 5.4.5. Redes Bayesianas

Una red Bayesiana (BN) [30,31] es un modelo gráfico que representa las relaciones de naturaleza probabilística entre las variables de interés. Este tipo de redes consiste en dos parte la primera cualitativa (modelo estructural), que proporciona una representación visual de las interacciones entre las variables, y una parte cuantitativa (a través de distribuciones locales de probabilidad ), que permite una inferencia probabilística y numéricamente mide el impacto de una variable fija o variable sobre otras. Tanto las partes cualitativos y cuantitativos determinar una distribución única de probabilidad conjunta sobre las variables de un problema específico [72, 73, 74]. En otras palabras, una red bayesiana es un grafo acíclico dirigido que consiste en [75]:

- Nodos (círculos), que representan variables aleatorias; arcos (flechas), que representan relaciones probabilísticas entre estas variables y
- para cada nodo, existe una distribución de probabilidad local conectado a ella, por lo que depende del estado de sus padres.

Una de las grandes ventajas de este modelo es que permite la representación de una distribución de probabilidad conjunta de una manera compacta

y económica haciendo uso extensivo de la independencia condicional, como se muestra en la ecuación 1:

$$P(X_1, X_2, ..., X_n) = \prod_{i=1}^{n} P(X_i|Pa(X_i)) \tag{5.10}$$

donde $Pa(X_i)$ representa el conjunto de nodos padres $X_i$; $i$.e.,nodos con arcos apuntando a $X_i$. Esta ecuación muestra como se obtiene una probabilidad conjunta, del producto de distribuciones locales de probabilidad condicional.

#### 5.4.5.1. Clasificadores Bayesianos

Los clasificadores Bayesianos utilizados en la presente tesis son: 1) Naïve Bayes, 2)Hill-Climber y c) Repeated-Hill Climber [11, 76, 77] serán descritos brevemente a continuación.

1. Naïve Bayes (NB) es uno de los clasificadores más eficaces [77]. Sus principales atractivos son su simplicidad y precisión, aunque su estructura es siempre fija (la variable de clase tiene un arco apuntando a cada atributo), se ha demostrado que este clasificador tiene una precisión de clasificación alta y un error de Bayes óptimo. En términos simples, el NB aprende, a partir de una muestra de datos de entrenamiento, de la probabilidad condicional de cada atributo dado a la clase. Entonces, una vez que llega un nuevo caso, el NB utiliza la regla de Bayes para calcular la probabilidad condicional de la clase dado el conjunto de atributos y selecciona el valor de la clase con la más alta probabilidad posterior.

2. Hill-Climber [11] es una implementación de un algoritmo de búsqueda y calificación en Weka , que utiliza greedy-hill climbing [78] para la parte de búsqueda y métricas diferentes para la parte de puntuación, como BIC, BD, AIC y MDL. Para los experimentos descritos aquí, hemos elegido la métrica MDL. Este procedimiento toma como entrada un gráfico vacío y una base de datos y se le aplican diferentes operadores para la construcción de una red bayesiana: adición, supresión o inversión de un arco. En cada etapa de búsqueda, busca una estructura que minimiza la puntuación MDL. En cada paso, el MDL se calcula y procedimiento Hill-Climber mantiene la estructura con el mejor (mínimo) puntuación. Termina la búsqueda cuando la nueva estructura no mejora la puntuación de MDL de la red anterior.

3. Repeated Hill-Climber [11] es la implementación de un algoritmo de búsqueda y de puntuación en Weka, que utiliza corridas repetidas de greedy-hill climbing [78] para la parte de búsqueda y métricas diferentes para la parte de puntuación, como BIC, BD, AIC y MDL. Para los

experimentos descritos aquí, hemos elegido la métrica MDL. En contraste con el sencillo algoritmo de Hill-Climber, Repeated Hill-Climber toma como entrada un gráfico generado aleatoriamente. También tiene una base de datos y aplica diferentes operadores (adición, supresión o inversión de un arco) y devuelve la mejor estructura de las corridas repetidas del procedimiento Hill-Climber. Con esta repetición de corridas, es posible reducir el problema de quedarse atascado en un mínimo local [74].

## 5.5. Evaluación de clasificadores

El método que ocupamos para la evaluación de los clasificadores es validación cruzada (cross-validation) de k-pliegues (k-fold). La definición de validación cruzada la tomamos de Kohavi [79]. En k-veces la validación cruzada divide el conjunto de datos $D$ en $K$ mutuamente excluyentes llamadas muestras aleatorias (folds): $D_1, D_2, \ldots, D_k$, donde los folds tienen aproximadamente el mismo tamaño. Se entrenaron a los clasificadores cada vez $i \in 1, 2, \ldots, k$ usando $D \setminus D_i$ y son probados en $Di$ (el símbolo $\setminus$ denota la diferencia entre conjuntos). La estimación de la validación cruzada es el número total de clasificaciones correctas, dividido por el tamaño de la muestra (número total de casos en $D$). Así la estimación de k-veces validación cruzada es:

$$acc_{cv} = \frac{1}{n} \sum_{(v_i, y_i) \in D} \delta(I(D \setminus D_{(i)}, v_i), y_i) \tag{5.11}$$

Donde $(I(D \setminus D_{(i)}, v_i), y_i)$ denota el nombre asignado por el clasificador $I$ a una instancia sin marcar $v_i$ en el conjunto de datos $D \setminus D_{(i)}$, $y_i$ es la instancia de la clase en $v_i$, $n$ es el tamaño completo del conjunto de datos y $\delta(i, j)$ es una función donde: $\delta(i, j) = 1$ si $i = j$ y 0 si $i \neq j$. Dicho de otra manera si la etiqueta asignada por el clasificador a la instancia sin marcar $v_i$ coincide con la clase $y_i$, entonces el resultado es de 1, de lo contrario, el resultado es 0; *i*.e., consideramos la pérdida 0/1 de la función en nuestros cálculos de la ecuación 7.1. Es importante mencionar que en la validación cruzada estratificada, los folds deben de contener (más o menos) la misma proporción de las clases como en el conjunto de datos completo $D$. Un caso especial de la validación cruzada se produce cuando $k = n$ (donde $n$ representa el tamaño de la muestra). Este caso se conoce como dejar uno fuera[80, 79].
Se evalúa el rendimiento de los clasificadores que se presentaron en las secciones previas,con las siguientes medidas [81, 82, 83, 84]:

*a*) *Precisión*: El número total de clasificaciones correctas dividido

por el tamaño del conjunto de prueba correspondiente. La fórmula es la siguiente.

$$p = \frac{tp}{n} \tag{5.12}$$

donde $tp$ es el número de casos clasificados correctamente y $n$ el número de casos de prueba.

*b*) *Sensibilidad*: Es la capicidad de identificar correctamente a aquellos pacientes que realmente tienen la enfermedad.

$$s = \frac{a}{a+c} \tag{5.13}$$

donde $a$ son los verdaderos positivos , $c$ son los falsos negativos y $a + c$ son el total de casos con la enfermedad.

*c*) *Especificidad*: Es la capacidad de identificar correctamente a aquellos pacientes que no tienen la enfermedad.

$$e = \frac{d}{b+d} \tag{5.14}$$

donde $d$ son los verdaderos negativos , $b$ son los falsos positivos y $b + d$ son el total de casos sin la enfermedad.

Tabla 5.3: Sensibilidad y especificidad de una prueba.

| **Resultado de la prueba** | **Sujetos con la enfermedad** | **Sujetos sin la enfermedad** |
|---|---|---|
| Positiva | Verdaderos positivos ($a$) | Falsos positivos ($b$) |
| Negativa | Falsos negativos ($c$) | Verdaderos negativos ($d$) |

# Capítulo 6

# Metodología y Resultados

## 6.1. Metodología

El conjunto de datos que se ocupa para los experimentos es el obtenido del análisis termográfico, conjunto definido en el capítulo 5.

Para realizar los experimentos de estadística descriptiva, sólo se utilizaron medidas descriptivas definidas en el capítulo 5, las cuales son media o promedio, mediana y desviación estándar.

Para realizar las correlaciones visuales se utilizó el software Weka [11], el estudio de correlación se le realizó a cada una de las variables con respecto a la clase. Para las correlaciones lineales con coeficiente de Pearson descrito en el capítulo 5, se utilizó el software estadístico SPSS [85]. Se llevó a cabo el coeficiente de Pearson para cada una de las variables de la termografía con respecto a la clase.

## 6.2. Resultados

### 6.2.1. Resultados de estadística descriptiva

En la tabla 6.1 se muestras las diferentes pruebas realizadas con distintos valores de score. Para fines de los experimentos aquí reportados el rango de score utilizado fue el de 160 (se eligió este valor porque era el que tenía mayor sensibilidad y especificidad), lo que quiere decir que para un estudio de termografía si la sumatoria de las variables termográficas es mayor a 160 es probable que exista una lesión cancerosa y si el score es menor a 160 es probable que la lesión sea benigna. Se contrastaron los resultados de los estudios termográficos con los obtenidos por la biopsia [86]. Los experimentos se realizaron con una base

de datos de 98 casos (descrita en el capítulo 4), de los cuales 77 son casos con lesiones cancerosas y 21 tienen lesiones benignas.

En la tabla 6.2 se presentan el número de aciertos y fallos, tanto para las lesiones benignas como para las lesiones malignas de los 98 casos con score de 160. En la tabla 6.3 a partir de este score se determina la precisión, sensibilidad y especificidad.

Tabla 6.1: Resultados de aciertos y fallos para casos con/sin cáncer utilizando como prueba de pre-diagnóstico la termografía, con diferentes valores de score

| | **Score 140** | **Score 150** | **Score 160** | **Score 170** | **Score 180** |
|---|---|---|---|---|---|
| Aciertos con cáncer | 64 | 60 | **54** | 51 | 38 |
| Fallos con cáncer | 13 | 17 | **23** | 26 | 39 |
| Aciertos sin cáncer | 3 | 8 | **10** | 11 | 15 |
| Fallos sin cáncer | 18 | 13 | **11** | 10 | 6 |

Tabla 6.2: Resultados de aciertos y fallos para casos con/sin cáncer utilizando como prueba de pre-diagnóstico la termografía, con score de 160

| **Enfermedad** | **Casos** | **Aciertos** | **Fallos** | **% aciertos** | **% fallos** |
|---|---|---|---|---|---|
| Lesiones con cáncer | 77 | 54 | 23 | 70.12 % | 29.88 % |
| Lesiones benignas | 21 | 10 | 11 | 47.62 % | 52.38 % |
| Total | 98 | 64 | 34 | 68.31 % | 34.69 % |

Tabla 6.3: Resultados de precisión, sensibilidad y especificidad para casos con/sin cáncer utilizando como prueba de pre-diagnóstico la termografía, con score de 160.

| | **Precisión** | **Sensibilidad** | **Especificidad** |
|---|---|---|---|
| Termografía mamaria | 68.31 % | 70 % (60-80) | 52 % (31-74) |

A partir de los estudios de patología se obtiene el tamaño de los tumores malignos, para fines de estos experimentos los tamaños del tumor se discretizaron así: 1 para tumores $< 2$ cm, 2 para tumores 2-5cm, 3 para

tumores > 5cm, en la tabla 6.4 se muestra la relación entre el tamaño del tumor y el score, para poder apreciar de manera visual la relación entre el tamaño del tumor y el score ver la figura 6.1.

Tabla 6.4: Resultados del score en relación a los tamaños del tumor.

| Tamaño | Promedio | Mediana | Moda | Desviación Estándar |
|---|---|---|---|---|
| 1 | 154.48 | 152 | 148 | 35.53 |
| 2 | 172.39 | 180 | 199 | 32.02 |
| 3 | 195.27 | 194.5 | 178 | 26.60 |

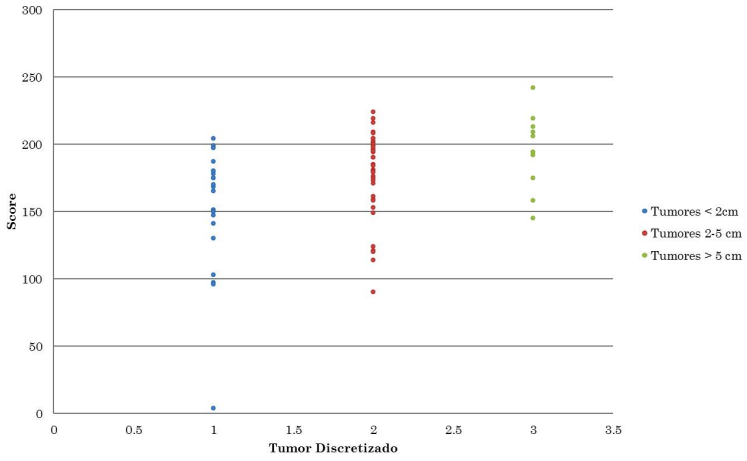


Figura 6.1: Relación entre el tamaño del tumor y el score.

Otro dato obtenido del estudio de patología es el tipo de lesión que presenta la mama. Para fines de este trabajo se consideraron 7 lesiones cancerosas y 8 lesiones benignas. En la tabla 6.5 se presentan los resultados de la relación entre el tipo de lesión(cancerosa, no-cancerosa) y el score obtenido.

De los 77 casos encontrados con cáncer de mama, el estudio de patología obtiene otro dato que es el grado de malignidad lo que significa que tan avanzada y agresiva está la enfermedad al momento de realizar la biopsia, a mayor grado es más agresiva la lesión. En la tabla 6.6 se presenta la relación entre el grado de la lesión cancerosa y el score.

Además de los datos termográficos de los pacientes, también se cuentan con sus respectivos estudios de mastografía, esta maneja un rango para dar un pre-diagnóstico. A este rango se le llama BIRADS y comienza en el 2, el cual indica que no existe una lesión, le sigue el 3 que indica que hay una lesión pero es poco probable que sea maligna, el 4 indica

Tabla 6.5: Resultados del score en relación al tipo de lesión, para los 98 casos. Las lesiones del 1-7 son cáncerosas y del 8-15 son lesiones no cancerosas

| No | Tipo Enfermedad | Casos | Promedio | Mediana | Desviación Estándar |
|---|---|---|---|---|---|
| 1 | CDI | 65 | 176.89 | 180 | 32.89 |
| 2 | CLI | 6 | 172 | 174 | 44.63 |
| 3 | CPMIC | 1 | 90 | 90 | |
| 4 | CMUC | 2 | 120.5 | 120.5 | 0.71 |
| 5 | CDIS | 2 | 194 | 194 | 2.83 |
| 6 | SARC | 0 | | | |
| 7 | Cmepid | 1 | 197 | 197 | |
| 8 | HDA | 5 | 163.20 | 152 | 24.30 |
| 9 | MST | 9 | 176.55 | 181 | 24.31 |
| 10 | MFQ | 0 | | | |
| 11 | FAM | 2 | 178 | 178 | 0 |
| 12 | ECT | 1 | 59 | 59 | |
| 13 | HDS | 1 | 130 | 130 | |
| 14 | PHY | 1 | 124 | 124 | |
| 15 | HFE | 1 | 148 | 148 | |

Tabla 6.6: Resultados del score en relación al grado de la lesión.

| Grado | Casos | Promedio | Mediana | Desviación Estándar |
|---|---|---|---|---|
| 1 | 4 | 148.75 | 150.5 | 52.44 |
| 2 | 38 | 173.37 | 179 | 30.87 |
| 3 | 36 | 178.78 | 191 | 36.63 |

que existe una lesión y es probable que sea maligna y el 5 indica que es muy probable que la lesión sea maligna. En la tabla 6.7, se muestra la relación entre el score y el valor de BIRADS para los pacientes diagnosticados con cáncer y en la tabla 6.8 para pacientes sin cáncer.

De las 16 variables que conforman el estudio de la termografía solo 8 variables (asimetría, red termovascular, patrón curvilíneo, porcentaje, hipertermía funcional, 2c, única F, Unilateral 1c, descritas en el capítulo 4) se suman para conformar el valor del score, una variable (GAP) es numérica y solo es informativa al oncólogo, de igual manera las otras 7 variables (axila, surco, centro caliente, forma irregular, pin point, histograma y perfil alterado) restantes toman 2 valores (positivo o negativo). Aunque no aportan peso al score, le ayudan al oncólogo en conjunción con el score a poder tomar una decisión. En la tabla 6.9, se

Tabla 6.7: Resultados del score en relación al valor de BIRADS para los 77 casos con cáncer.

| **BIRADS** | **Frecuencia** | **Promedio** | **Mediana** | **Desviación Estándar** |
|---|---|---|---|---|
| 2 | 5 | 182.80 | 197 | 28.11 |
| 3 | 3 | 196.33 | 195 | 7.09 |
| 4 | 21 | 170.71 | 180 | 37.73 |
| 5 | 48 | 174.12 | 180 | 35.73 |

Tabla 6.8: Resultados del score en relación al valor de BIRADS para los 21 casos sin cáncer.

| **BIRADS** | **Frecuencia** | **Promedio** | **Mediana** | **Desviación Estándar** |
|---|---|---|---|---|
| 2 | 1 | 145 | 145 | |
| 3 | 5 | 187 | 178 | 18.87 |
| 4 | 14 | 156.14 | 151.50 | 36.56 |
| 5 | 1 | 130 | 130 | |

muestra la relación entre el score y los valores que toman las variables, tanto para casos positivos como negativos para el carcinoma ductal infiltrante dado que es el tipo de cáncer de mama que se presentó con mayor frecuencia en este conjunto de datos. Además, en la tabla 6.10 se muestra la relación entre las 7 variables y los 21 casos sin cáncer.

Tabla 6.9: Resultados del score en relación con las 7 variables que toman valores de positivo y negativo para el carcinoma ductal infiltrante (66 de los 77 casos con cáncer).

| **Variable** | **Positivos** | **Negativos** | **Score Positivos** | **Score Negativos** |
|---|---|---|---|---|
| Axila | 37 | 29 | 178.60 | 178.70 |
| Surco | 54 | 12 | 182.15 | 164.17 |
| Centro caliente | 66 | 0 | 178.88 | 0 |
| Forma irregular | 64 | 2 | 180 | 143 |
| Pin point | 59 | 7 | 179.90 | 170.28 |
| Histograma | 58 | 8 | 181.86 | 157.25 |
| Perfil alterado | 31 | 35 | 186 | 170.84 |

Tabla 6.10: Resultados del score en relación con las 7 variables que toman valores de positivo y negativo para los 21 casos sin cáncer.

| Variable | Positivos | Negativos | Score Positivos | Score Negativos |
|---|---|---|---|---|
| Axila | 10 | 11 | 146.40 | 173.82 |
| Surco | 13 | 8 | 167.15 | 150.38 |
| Centro caliente | 19 | 2 | 162.26 | 146.50 |
| Forma irregular | 17 | 4 | 158.53 | 170.25 |
| Pin point | 14 | 7 | 170.50 | 141.29 |
| Histograma | 14 | 7 | 157.57 | 167.14 |
| Perfil alterado | 6 | 15 | 181 | 152.67 |

### 6.2.2. Resultados de correlaciones

En la figuras 6.2- 6.17, se muestran las correlaciones visuales entre las 16 varibles termográficas y la clase (Enfermo, No_Enfermo), las cuales se obtuvieron mediante el software Weka [11]. En la tabla 6.11 se presentan los resultados del coeficiente de correlación lineal de Pearson, entre las variables termográficas y la clase, dichas correlaciones se obtuvieron mediante el software estadístico SPSS [85].

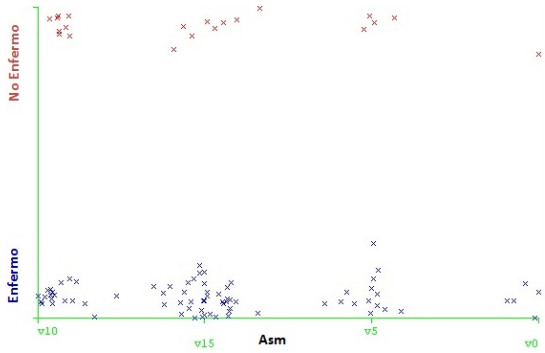


Figura 6.2: Correlación visual entre la variable Asimetría y la clase (Enfermo, No_Enfermo).

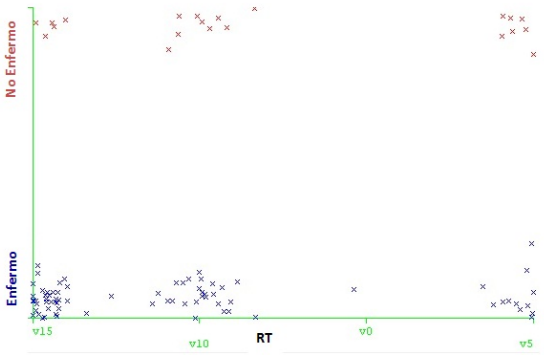


Figura 6.3: Correlación visual entre la variable Red Termovascular y la clase (Enfermo, No_Enfermo).

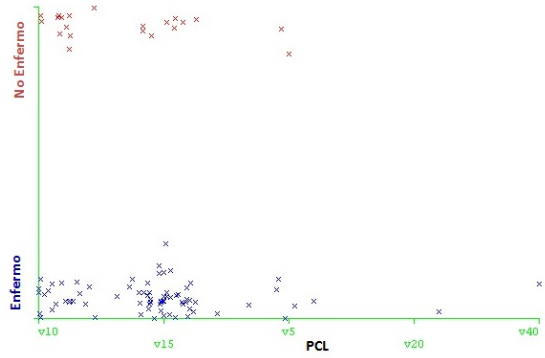


Figura 6.4: Correlación visual entre la variable Patrón curvilíneo y la clase (Enfermo, No_Enfermo).

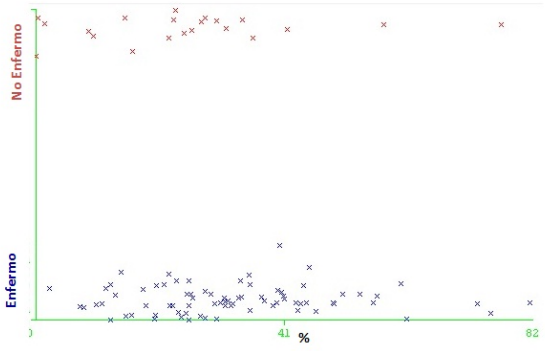


Figura 6.5: Correlación visual entre la variable Porcentaje y la clase (Enfermo, No_Enfermo).

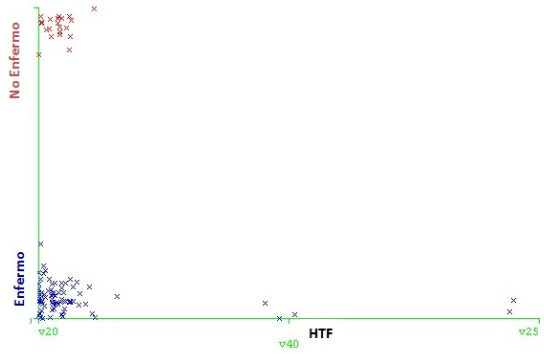


Figura 6.6: Correlación visual entre la variable Hipertermía y la clase (Enfermo, No_Enfermo).

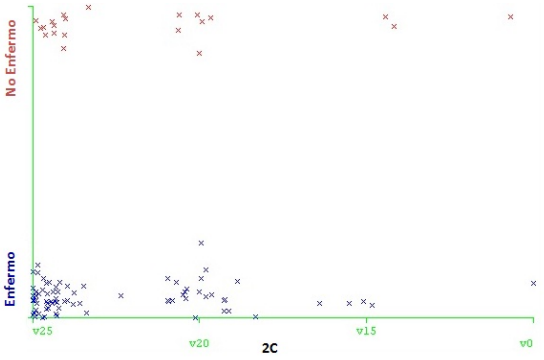


Figura 6.7: Correlación visual entre la variable Diferencia entre cimas y la clase (Enfermo, No_Enfermo).

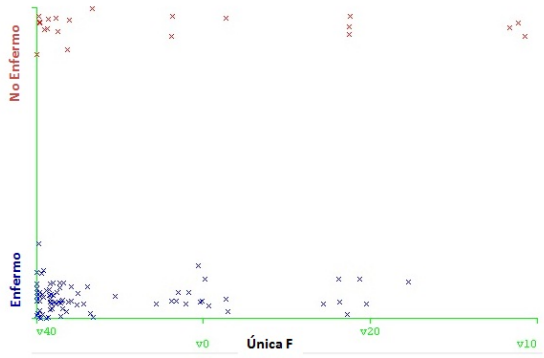


Figura 6.8: Correlación visual entre la variable Única cima y la clase (Enfermo, No_Enfermo).

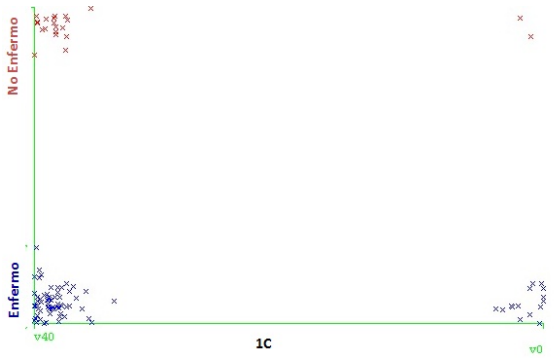


Figura 6.9: Correlación visual entre la variable Unilateral y la clase (Enfermo, No_Enfermo).

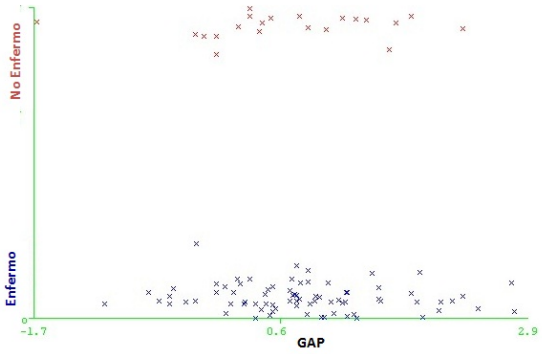


Figura 6.10: Correlación visual entre la variable GAP y la clase (Enfermo, No_Enfermo).

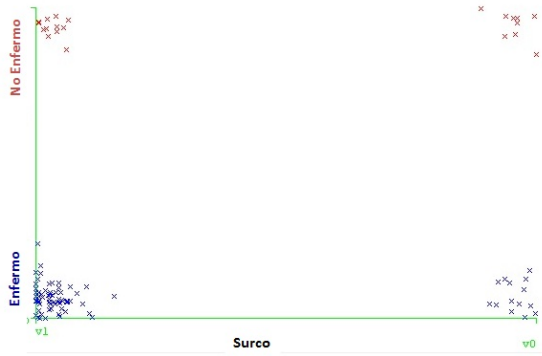


Figura 6.11: Correlación visual entre la variable Surco y la clase (Enfermo, No_Enfermo).

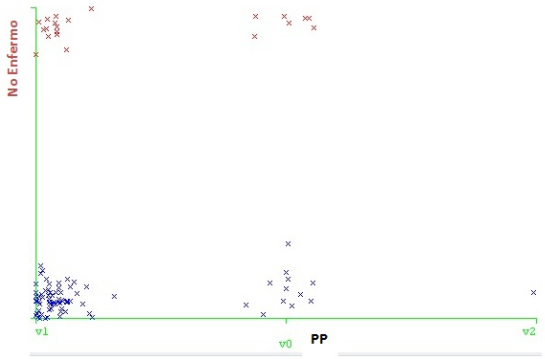


Figura 6.12: Correlación visual entre la variable Pin point y la clase (Enfermo, No_Enfermo).

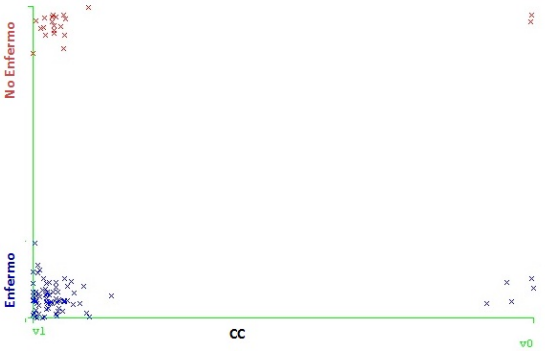


Figura 6.13: Correlación visual entre la variable Centro caliente y la clase (Enfermo, No_Enfermo).


Figura 6.14: Correlación visual entre la variable Forma irregular y la clase (Enfermo, No_Enfermo).

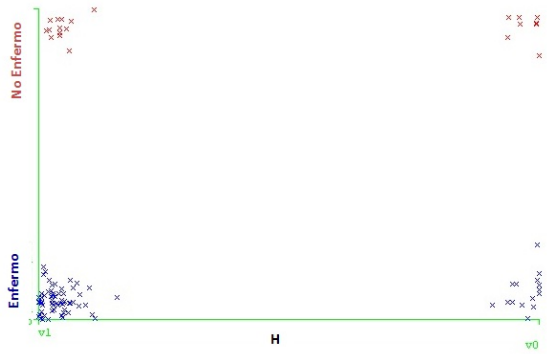


Figura 6.15: Correlación visual entre la variable Histograma y la clase (Enfermo, No_Enfermo).


Figura 6.16: Correlación visual entre la variable Axila y la clase (Enfermo, No_Enfermo).

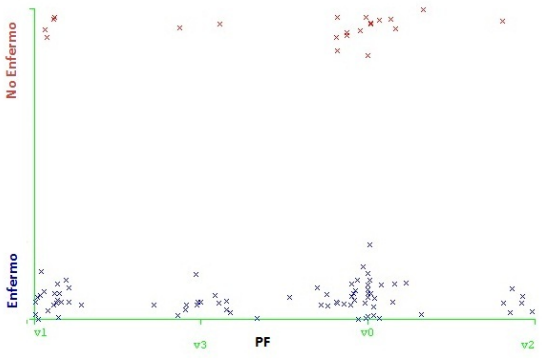


Figura 6.17: Correlación visual entre la variable Perfil Alterado y la clase (Enfermo, No_Enfermo).

Tabla 6.11: Resultados de la correlación entre las 16 variables termográficas y la clase (Enfermo, No_Enfermo).

| No | Variable | Coeficiente de Pearson | S. Bilateral |
|---|---|---|---|
| 1 | Asimetría | .036 | .728 |
| 2 | Red termovascular | .185 | .075 |
| 3 | Patrón curvilíneo | .135 | .193 |
| 4 | Porcentaje | .186 | .073 |
| 5 | Hipertermía | .112 | .283 |
| 6 | Diferencia entre cimas | .140 | .177 |
| 7 | Única cima | .111 | .285 |
| 8 | Unilateral | .208 | -.131 |
| 9 | GAP | -.015 | .889 |
| 10 | Surco | .186 | .073 |
| 11 | Pin point | .193 | .062 |
| 12 | Centro caliente | .042 | .685 |
| 13 | Forma irregular | -.036 | .734 |
| 14 | Histograma | .197 | .057 |
| 15 | Axila | .194 | .061 |
| 16 | Perfil Alterado | .159 | .125 |

### 6.2.3. Resultados de los clasificadores

Para el conjunto de datos de la termografía fueron ejecutados diferentes clasificadores. En la Tabla 6.12 se muestran los resultados de la ejecución en Weka de 3 clasificadores Bayesianos, junto a la precisión se muestra la desviación estándar. Para el resto de las pruebas les corresponden un intervalo de confianza (IC) del 95 % indicado entre paréntesis y en las Tablas 6.13- 6.15 están sus respectivas matrices de confusión (es una matriz que indica el número de casos clasificados correctamente, los casos en los que se equivoco y a que clase corresponden). En la Tabla 6.16 se muestran los resultados de la ejecución de los clasificadores: K-NN, AdaBoost, árboles de decisión ID3 y C4.5. En las Tablas 6.17- 6.20 se muestran las matrices de confusión para estos clasificadores. En las figuras 6.13 y 6.15, se muestran las redes bayesianas Hill-Climber y Repeated Hill-Climber respectivamente. En la figura 6.20 se puede observar el árbol generado por el clasificador C4.5.

Tabla 6.12: Precisión, sensibilidad y especificidad de las redes bayesianas, para la termografía mamaria.

| | **Naïve Bayes** | **Hill-Climber** | **Repeated Hill-Climber** |
|---|---|---|---|
| Precisión | 70.19 % (±12,24) | 78.56 % (±3,0) | 78.56 % (±3,0) |
| Sensibilidad | 82 % (73-90) | 100 % (100-100) | 100 % (100-100) |
| Especificidad | 33 % (13-53) | 0 % (0-0) | 0 % (0-0) |

Tabla 6.13: Matriz de confusión de Naïve Bayes.

| **Enfermo** | **No-Enfermo** | **Clasificado como** |
|---|---|---|
| 63 | 14 | Enfermo |
| 14 | 7 | No-Enfermo |

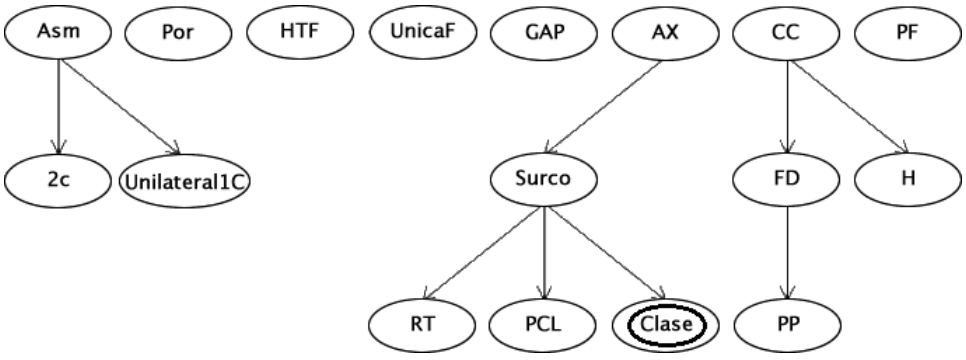


Figura 6.18: Red Bayesiana con procedimiento Hill-Climber generada de la base de datos de la termografía.

Tabla 6.14: Matriz de confusión de Hill-Climber.

| Enfermo | No-Enfermo | Clasificado como |
|---|---|---|
| 77 | 0 | Enfermo |
| 21 | 0 | No-Enfermo |

Tabla 6.15: Matrix de confusión de Repeated Hill-Climber.

| Enfermo | No-Enfermo | Clasificado como |
|---|---|---|
| 77 | 0 | Enfermo |
| 21 | 0 | No-Enfermo |

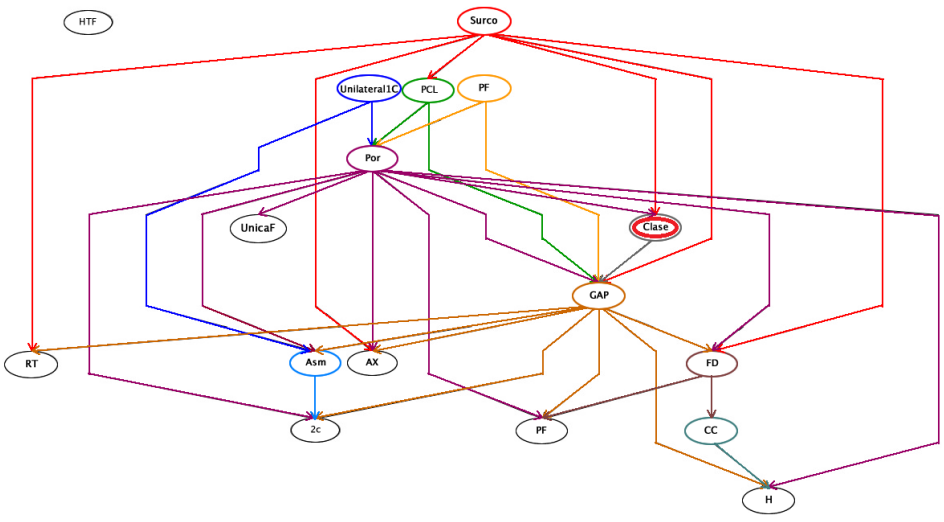


Figura 6.19: Red Bayesiana con procedimiento Repeated Hill-Climber generada de la base de datos de la termografía.

Tabla 6.16: Precisión, sensibilidad y especificidad de los clasificadores: KNN, AdaBoost, arboles de decisión ID3 y C4.5, para las 16 variables de termografía mamaria

| | **K-NN** | **AdaBoost** | **ID3** | **C4.5** |
|---|---|---|---|---|
| Precisión | 75.92 % (±12,50) | 77.52 % (±6,51) | 74.06 % (±12,12) | 75.68 % (±9,93) |
| Sensibilidad | 87 % (80-95) | 97 % (94-100) | 87 % (80-95) | 95 % (90-100) |
| Especificidad | 33 % (13-53) | 10 % (-3-22) | 48 % (26-69) | 10 % (-3-22) |

Tabla 6.17: Matriz de confusión de K-NN.

| **Enfermo** | **No-Enfermo** | **Clasificado como** |
|---|---|---|
| 67 | 10 | Enfermo |
| 14 | 7 | No-Enfermo |

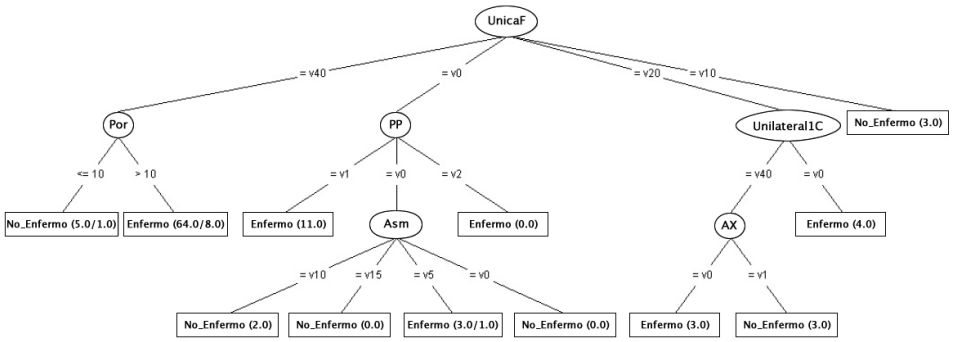


Figura 6.20: Árbol de decisión C4.5 generado de la base de datos de la termografía.

Tabla 6.18: Matriz de confusión de AdaBoost.

| **Enfermo** | **No-Enfermo** | **Clasificado como** |
|---|---|---|
| 75 | 2 | Enfermo |
| 19 | 2 | No-Enfermo |

Tabla 6.19: Matriz de confusión de ID3.

| **Enfermo** | **No-Enfermo** | **Clasificado como** |
|---|---|---|
| 65 | 9 | Enfermo |
| 11 | 10 | No-Enfermo |

Tabla 6.20: Matriz de confusión de C4.5.

| **Enfermo** | **No-Enfermo** | **Clasificado como** |
|---|---|---|
| 73 | 4 | Enfermo |
| 19 | 2 | No-Enfermo |

# Capítulo 7

# Discusión

El estudio termográfico mamario es una técnica de pre-diagnóstico novedosa y de reciente utilización en México. Por lo tanto en este trabajo se realizaron pruebas con diferentes rangos de score como se puede observar en la tabla 6.2 para determinar cuál sería el umbral que tendría el score, para así poder determinar a partir del score si una lesión es maligna o benigna, se encontró que el umbral del score que mejor funcionaba era el de 160, ya que este tenía un mejor equilibrio entre la sensibilidad y la especificidad, como se puede observar en las tablas 6.2 y 6.3.

Una vez que se fijó el punto de corte del score, se obtuvieron medidas descriptivas para poder observar a grandes rasgos el comportamiento del conjunto de datos termográficos. Entre los primeros experimentos se buscó una relación entre el tamaño del tumor obtenido de la biopsia y el score (sólo para los tumores con cáncer), dicha relación se puede ver en la figura 6.1 donde efectivamente a mayor tamaño del tumor el valor del score aumenta, sin embargo sí se contrasta con la tabla 6.4 donde aparece además de la media del score la desviación estándar para los tres tamaños de tumores, la desviación estándar es muy grande lo que sugiere que hay mucho ruido en los datos. Otro dato obtenido de la biopsia sólo para los casos con cáncer es el grado de malignidad que presenta el tumor al realizar la biopsia, en su mayoría los tumores están en grado 2 y 3 ésto se puede observar en la tabla 6.6, lo que nos indica una detección tardía de la enfermedad. Pareciera haber una relación entre el grado de malignidad y el score, a mayor grado le corresponde un score más alto, pero existe una desviación estándar alta.

Por otro lado, se buscó la relación existente el score y las distintas enfermedades que detecta la biopsia. Ésta puede detectar 15 tipos de

enfermedades, de las cuales 7 tipos son lesiones malignas y 8 son lesiones benignas. La relación observada en la tabla 6.5 es muy poco informativa, ya que la muestra con la que se cuenta es muy pequeña (98 casos), aunque tenemos para los diferentes tipos de enfermedad en su mayoría al menos un caso, no se puede determinar si el comportamiento del score tiene un patrón, pero por ejemplo para el tipo de enfermedad 1 Çarcinoma Ductal Infiltrante", es la enfermedad con más casos del conjunto de datos, ésta posee en promedio un score de 176.89, lo que nos haría pensar que el score de 160 podría determinar un buen número de este tipo de enfermedad, pero lo que luce desalentador es la desviación estándar tan grande de 32.89 lo que indica que hay una dispersión muy grande de los datos de la media.

Aunado a los datos de patología de los pacientes, también se tienen los estudios da mastografía. Como se mencionó previamente, ésta se mide con BIRADS, es un tipo de puntaje que va de 2-5, si el paciente tiene como resultado BIRADS = 2 ó 3 es poco probable que presente una lesión cancerosa y a consideración del médico tratante se le puede o no practicar una biopsa, pero si BIRADS se encuentra entre 4 y 5 es muy probable que presente una lesión cancerosa y por protocolo se le debe realizar una biopsia, por lo tanto a mayor BIRADS mayor probabilidad de que el paciente tenga una lesión cancerosa. En las tablas 6.7 y 6.8 podemos observar la relación entre el puntaje dado por la mastografía y el score de la termografía. En la tabla 6.7 para este experimento fueron tomados los 77 casos con cáncer de los cuales 8 de ellos presentaron una puntuación baja (BIRADS 2 y 3) en la mastografía lo que podría causar una baja probabilidad de cáncer, sin embargo la termografía obtuvo un promedios altos 182.80 y 196.33, pero se presentan desviaciones estándar grandes (28.11 y 7.09), pero si la mastografía trabaja en conjunto con la termografía en estos casos sería más probable que el médico tratante puede ubicarlos, que en condiciones normales, sería probable que se dejaran pasar. En la tablan 6.8 fueron utilizados los 21 casos sin cáncer de los cuales 15 tuvieron un BIRADS entre 4 y 5, por lo tanto por protocolo se les realizó una biopsia, pero en el estudio termográfico obtuvieron una puntuación menor a 160, aunque la biopsia se les tuvo que realizar, el médico tratante con la información termográfica podría tranquilizar al paciente al mencionarle que es poco probable que sea cáncer y si lo es se encuentra en una etapa temprana.

Hay 7 variables que no aportan peso al score, pero influyen en la decisión que el oncólogo toma para dar un pre-diagnóstico, estas variables toman el valor de positivo o negativo. Si observamos el comportamiento de estas variables para los 66 casos con carcinoma ductal infiltrante

en la tabla 6.9, la variable Axila tiene 39 positivos y 27 negativos, está variable no es un buen discriminante porque a partir de ella no se puede dar un pre-diagnóstico de cáncer dado que la probabilidad para este conjunto de datos es del 59 %. Otra variable que presenta un comportamiento parecido al anterior es la de perfil alterado con 31 positivos y 35 negativo. Por otro lado el panorama lucía alentador con la variables: centro caliente con 66 Positivos y 0 negativos, o la variable forma irregular con 64 positivos y 2 negativos y surco con 54 positivos y 12 negativos, se podría concluir que a partir de estas variables sería posible discriminar para dar un pre-diagnóstico con alta probabilidad de cáncer, pero si se observa la tabla 6.10 para los 21 casos sin cáncer, estas variables también se comportan así, centro caliente 19 positivos y 2 negativos, forma irregular 17 positivos y 4 negativos. De ahí que se puede concluir que estos 21 casos sin cáncer, tendrían alta probabilidad de tener cáncer cuando ésto no es cierto.

Como ya se obtuvieron valores estadísticos y se observó que existía ruido, en las figuras 6.2- 6.17 se observa la relación que existe entre cada una de las variables termográficas y la clase (Enfermo, No_Enfermo). Como se puede observar, en todas las variables termográficas sus valores se traslapan con respecto a la clase, con ésto nos referimos a que los valores de las variables termográficas son utilizados tanto para la clase enfermo como para la No_Enfermo. Esto nos lleva a pensar que hay una correlación mínima o nula entre estas variables y la clase. Este supuesto se confirma con la tabla 6.11 en donde se obtiene el coeficiente de correlación lineal de Pearson, el cual entre más cercano se encuentre a 0 la correlación entre la clase y las variables termográficas es menor, la variable que tiene un mayor coeficiente es la unilateral (.208), le sigue surco (.186), histograma (.197), axila(.194) y las más bajas son: Gap (-.015) éste mínima correlación es inversa, asimetría(.36) y centro caliente (.042).

Con los resultados de la correlación, se puede esperar que los clasificadores presenten un bajo desempeño. Los clasificadores con los que se probó fue con los de Redes Bayesianas: Naïve Bayes, Hill-Climber y Repeated Hill-Climber, cuyos resultados de precisión, sensibilidad, especificidad y las matrices de confusión, se encuentran en las tablas 6.12- 6.15. Otros clasificadores con los que se probó son: K-NN, AdaBoost, árboles de decisión ID3 y C4.5. Los resultados de éstos se pueden consultar en la tabla 6.16 y sus respectivas matrices de confusión en las tablas 6.17- 6.20. El clasificador del grupo previamente mencionado que obtuvo un mejor porcentaje de clasificación fue AdaBoost con 77.52 %, pero Hill-Climber y Repeated Hill-Climber, presentaron

un mayor porcentaje de clasificación con un 78.56 % y son muy buenos para predecir cuándo se tiene la enfermedad o dicho de otra forma tienen una sensibilidad del 100 %, pero una especificidad nula con un 0 %, todo ello sugiere que el clasificador no sabe distinguir a los pacientes sin la enfermedad. El clasificador que presenta mejor especificidad es ID3 con un 48 %. Un patrón presente en las matrices de confusión fue que los clasificadores se equivocaron mayormente al clasificar los casos que tenían la clase No_Enfermo. Otro dato interesante de los clasificadores Hill-Climber y Repeated Hill-Climber, son las redes obtenidas (ver figuras 6.18 y 6.19) donde donde se puede apreciar cuáles son las variables que están directamente conectadas con la clase; para la red obtenida de Hill-Climber (ver figura 6.18) la única variable que está directamente conectada es el surco, por lo tanto podríamos quitar las otras variables y el porcentaje de precisión sería muy cercano al 78.56 % sólo con esta variable, sin tener las otras 15 variables termográficas. Para la red obtenida a partir de Repeated Hill-Climber (ver figura 6.19) se presentó la misma situación que la red anterior, sólo esta directamente relacionada con la variable Surco. Por otro lado, se generó el árbol de decisión C4.5 (ver figura 6.20), como se puede apreciar, el árbol se formo con sólo 6 variables, con ésto se podría concluir que con éstas variables obtendrían un porcentaje de clasificación cercano al obtenido con las 16 variables.

# Capítulo 8

# Conclusiones y trabajos futuros

La termografía mamaria es una técnica complementaria que ha sido retomada en los últimos años porque se han demostrado sus potencialidades como auxiliar a las técnicas de pre-diagnóstico ya existentes. Actualmente en México este estudio sólo es realizado en una clínica de Oaxaca y en Xalapa por un oncólogo quien fue el que proporciono los datos termográgicos que fueron utilizados para realizar los experimentos de la presente tesis. Derivado el presente trabajo de tesis se obtuvo una clasificación de artículos relacionados con la termografía, la cual no se había reportado en la literatura especializada.

También se describió un procedimiento para la toma y el análisis de las imágenes infrarrojas de la termografía mamaria, así como la formación del score, describiendo cómo es obtenida cada una de las variables termográficas.

Retomando la hipótesis realizada al principio de este trabajo .Existe consistencia entre la información cuantitativa del conjunto de datos termográficos y el pre-diagnóstico dado por el oncólogo", dados los experimentos realizados con el conjunto de datos termográficos con 98 casos, no existe como tal una consistencia entre los datos y el pre-diagnóstico que emite el oncólogo, debido a la cantidad de ruido que presentan los datos. Ésto se puede observar en las desviaciones estándar grandes en los cálculos de las medias a partir del score, también ésto se debe a la baja correlación existente entre las variables termográficas y la clase. Los experimentos se vieron limitados por el número de casos, también influyó el sesgo de los datos, ya que hay una clase con más casos (Enfermo) y otra clase con menores ejemplos (No_enfermo), éste sesgo

esta dado porque los datos fueron recopilados por el oncólogo de un hospital de cancerología, cuando los pacientes son derivados a está clínica es porque es muy alta la probabilidad de tener cáncer. Por último, el procedimiento que utiliza desde su expertiz el oncólogo para poder emitir un pre-diagnóstico no está proporcionando toda la información que ocupa para poder dar un pre-diagnóstico semi-automatizado.

En lo que respecta a los experimentos realizados con los clasificadores se puede llegar a la conclusión que la termografía puede detectar casos con cáncer ya que posee una buena sensibilidad donde algunos clasificadores alcanzan hasta el 100 %, los clasificadores que logran obtener este resultado son: Hill-Climber y Repeated Hill-Climber, pero les va muy mal al tratar de clasificar los casos sin la enfermedad, teniendo una especificidad nula del 0 %, en la mayoría de los clasificadores la especificidad es muy baja. El clasificador que obtuvo mejor especificidad fue ID3 con un 48 %. Otros valores que nos aportan información de la situación del conjunto de datos son las matrices de confusión, en ellas se aprecian que las equivocaciones de los clasificadores en su mayoría son en la clase de No_Enfermo, lo que lleva a una baja especificida. Dado los resultados anteriores la termografía podría ser utilizada como herramienta complementaria a las técnicas de pre-diagnóstico. Una de las ventajas de la termografía son los bajos costos empleados en la realización del estudio y la portabilidad para poder llegar a lugares inaccesibles para otras técnicas de pre-diagnóstico. Podría ser utilizada como método de pre-diagnóstico, debido que si indica que se tiene cáncer es muy probable que eso sea dado a su alta sensibilidad.

Se puede entonces concluir que no son necesarias todas las variables de la termografía para poder dar un pre-diagnóstico, ya que no todas las variables están directamente conectadas con la variable clase, como se puede observar en las redes Bayesianas presentadas en el capítulo 6; ésto es confirmado con el árbol de decisión C4.5 construido con sólo 6 variables, presentado en el mismo capítulo.

Como trabajo futuro, dado que el conjunto de datos de la termografía esta más sesgados hacia una clase (Enfermo), sería interesante ver el desempeño de los clasificadores con las clases balanceadas [87]. Como otra continuación a este trabajo, se propone encontrar una función más compleja que no sólo sea una sumatoria de variables como se tiene hasta ahora para calcular el score,y con ello así mejorar la sensibilidad y la especificidad de la termografía. Otra línea que queda abierta, es el poder reducir la subjetividad que se encuentra implícita en la mayoría de las variables termográficas. Para finalizar, como se observó en este

trabajo de tesis no son necesarias todas las variables termográficas, dado que con un número menor de variables se tiene una precisión de clasificación muy cercana a la que se tiene con todas ellas, se podría buscar la reducción del número de variables.

# Referencias


[1] Knaul F.M., Nigenda G., Lozano R., Arreola O.H., Langer A., and Frenk J. Breast cancer in mexico: a pressing priority. *Journal of Midwifery & Women's Health*, 16:113–123, 2008.

[2] Nosarti C., Roberts J.V., Crayford T., McKenzie K., and David A.S. Early psychological adjustment in breast cancer patients: A prospective study. *Journal of Psychosomatic Research*, 53(6):1123 – 1130, 2002.

[3] Bonnema J., Van Geel A.N., Van Ooijen B., Mali S.P.M., Tjiam S.L., Henzen-Logmans S.C., Schmitz P.I.M, and Wiggers T. Ultrasound-guided aspiration biopsy for detection of nonpalpable axillary node metastases in breast cancer patients: New diagnostic method. *World Journal of Surgery*, 21:270–274, 1997.

[4] Schnall M.D., Blume J., Bluemke D.A., DeAngelis G.A., DeBruhl N., Harms S., Heywang-Köbrunner S.H., Hylton N., Kuhl C., Pisano E.D., Causer P., Schnitt S.J., Smazal S.F., Stelling C.B., Lehman C., Weatherall P.T, and Gatsonis C.A. Mri detection of distinct incidental cancer in women with primary breast cancer studied in ibmc 6883. *Journal of Surgical Oncology*, 92(1):32–38, 2005.

[5] Brandan M.E. and Villaseñor N.Y. Detección del cáncer de mama: Estado de la mamografía en méxico. *Revista del Instituto Nacional de Cancerología*, pages 147–162, 2006.

[6] Geller B.M., Kerlikowske K.C., Carney P.A., Abraham L.A., Yankaskas B.C., Taplin S.H., Ballard-Barbash R., Dignan M.B., Rosenberg R., Urban N., and Barlow W.E. Mammography surveillance following breast cancer. *Breast Cancer Research and Treatment*, 81:107–115, 2003.

[7] Arora N., Martins D., Ruggerio D., Tousimis E., Swistel A.J., Osborne M.P., and Simmons R.M. Effectiveness of a noninvasive

digital infrared thermal imaging system in the detection of breast cancer. *The American Journal of Surgery*, 196:523–526, 2008.

[8] Ng E.Y.K. A review of thermography as promising non-invasive detection modality for breast tumor. *International Journal of Thermal Sciences*, 48:849–859, 2009.

[9] EtehadTavakol M., Sadri S., and Ng E.Y.K. Application of k- and fuzzy c-means for color segmentation of thermal infrared breast images. *Journal of Medical Systems*, 34:35–42, 2010.

[10] Mitchell T.M. *Machine Learning*. MIT Press and The McGrawHill Companies,Inc., 2003.

[11] Witten I. H. and Frank E. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Series in Data Management Sys. Morgan Kaufmann, second edition, 2005.

[12] Tellez V.A. Extracción de información con algoritmos de clasificación. Master's thesis, Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE), 2005.

[13] Malaga Vascular Institute. Radiología vascular intervencionista. Consultado el 7/11/2011 de `http://www.malagavascular.com/es/para-pacientes/List/show/biopsia-de-mama-199`, 2009.

[14] EtehadTavakol M., Lucas C., Sadri S., and Ng E.Y.K. Analysis of breast thermography using fractal dimension to establish possible difference between malignant and benign patterns. *Journal of Healthcare Engineering*, 1:27–44, 2010.

[15] Hairong Q., Phani T.K., and Zhongqi L. Early detection of breast cancer using thermal texture maps. In *Biomedical Imaging, 2002. Proceedings. 2002 IEEE International Symposium on*, pages 309–312, 2002.

[16] Ng E.Y.K., Chen Y., and Ung L. N. Computerized breast thermography: study of image segmentation and tempe rature cyclic variations. *Journal of Medical Engineering Technology*, 25:12–16, 2001.

[17] Jemal A., Bray F., Center M., Ferlay J., Ward E., and Forman D. Global cancer statistics. *CA: A Cancer Journal for Clinicians*, 61(1):69–90, 2011.

[18] Martínez-Montañez O.G., Uribe-Zúñiga P., and Hernández-Ávila M. Políticas públicas para la detección del cáncer de mama en méxico. *Salud Pública de México*, 51(2):350–360, 2009.

[19] Gutierrez F., Vazquez J., Venegas L., Terrazas S., Marcial S., Guzman C., Perez J., and Saldana M. Feasibility of thermal infrared imaging screening for breast cancer in rural communities of southern mexico: The experience of the centro de estudios y prevencion del cancer (ceprec). In *2009 ASCO Annual Meeting*, page 1521. American Society of Clinical Oncology, 2009.

[20] Lostao L. *Detección Precoz del cáncer de mama: Factores asociados a la participación en un progrma de screening.* Díaz de Santos, 2001.

[21] Yong G., You-Quan C., Zu-Long C., Yuan-Gui G., Ning-Yu A., Lin M., Srikanth M., and Jia-Hong G. Differentiation of clinically benign and malignant breast lesions using diffusion-weighted imaging. *Journal of Magnetic Resonance Imaging*, 16(2):172–178, 2002.

[22] Bayo C. J., García M.J., and Lluch H.A. *Cáncer de mama Cuestiones más frecuentes.* Grupo Editorial Entheos, 2007.

[23] Jimenéz M. *Oncología: Cáncer de mama.* Arán Ediciones, S.L., second edition, 2009.

[24] Ryke Geerd Hamer. Nueva medicina germánica. Consultado el 1/11/2011 de `http://learninggnm.com/documents/terceraley.html`, 2011.

[25] Zepeda-Castilla E., JoséRecinos-Money E., Cuéllar-Hubbe M., Robles-Vidal C.D., and Maafs-Molina E. Clasificación molecular del cáncer de mama. *Cirugía y Cirujanos*, 76:87–93, 2008.

[26] Martin D. *Patología Quirúrgica.* Elsevier, 2005.

[27] Winchester D.P and Winchester D.J. *Cáncer de Mama Atlas de Oncología Clínica.* Elsevier España, 2001.

[28] Geller B.M., Kerlikowske K.C., Carney P.A., Abraham L.A., Yankaskas B.C., Taplin S.H., Ballard-Barbash R., Dignan M.B., Rosenberg R., Urban N., and Barlow W.E. Mammography surveillance following breast cancer. *Breast Cancer Research and Treatment*, 81:107–115, 2003.

[29] Centro de diagnóstico mamario la viña. Tips de salud. Consultado el 7/11/2011 de `http://www.cpv.com.ve/tips_medicos.htm`, 2010.

[30] Sierra G., Piñero M., and Illana M. *Cirugía de la mama.* Arán Ediciones, S.L., 2006.

[31] Foster K.R. Thermographic detection of breast cancer. *IEEE Engineering in Medicine and Biology Magazine*, 17:10–14, 1998.

[32] Wishart G.C., Campisi M., Boswell M., Chapman D., Shackleton V., Iddles S., Hallett A., and Britton P.D. The accuracy of digital infrared imaging for breast cancer detection in women undergoing breast biopsy. *European Journal of Surgical Oncology (EJSO)*, 36:535–540, 2010.

[33] Richter J.A., García-Velloso M.J., Domínguez I., Quincoces G., Prieto E., and Rodríguez F.M. Tomografía por emisión de positrones en el cáncer de mama. *Revista de medicina*, 52(1):4–12, 2008.

[34] Alvarez G. and Tejerina G. *Mastología Dinámica*. Ediciones Díaz Santos, S.A., 1995.

[35] International Academy of Clinical Thermology. What is breast thermography. Consultado el 7/11/2011 de `http://www.iact-org.org/patients/breastthermography/what-is-breast-therm.html`, 2010.

[36] J Anderson. Thermography: A holistic approach to breast screening. *Womens health*, pages 12–13, 2010.

[37] Ng E.Y.K. and Fok S.-C. A framework for early discovery of breast tumor using thermography with artificial neural network. *The Breast Journal*, 9:341–343, 2003.

[38] Sterns E.E. and Zee B. Thermography as a predictor of prognosis in cancer of the breast. *Cancer*, 67:1678–1680, 1991.

[39] Ohsumi S., Takashima S., Aogi K., and Usuki H. Prognostic value of thermographical findings in patients with primary breast cancer. *Breast Cancer Research and Treatment*, 74:213–220, 2002.

[40] Z. Damnjanovic, D. Petrovic, R. Pantovic, and Z. Smiljanic. Infra red digital imaging in medicine. *International Journal of Collaborative Research on Internal Medicine & Public Health*, pages 425–434, 2010.

[41] Isard H. J. Thermography in mass screening of cancer success and failures. *Annais New York Academic Science*, 335:489–491, 1980.

[42] Ng E.Y.K., Ung L.N., Ng F.C., and Sim L.S.J. Statistical analysis of healthy and malignant breast thermography. *Journal of Medical Engineering Technology*, 25:253–263, 2001.

[43] Draper J.W. and Boag J.W. The calculation of skin temperature distributions in thermography, physics in medicine and biology. *Physics in Medicine and Biology*, 16:201–211, 1971.

[44] Ng E.Y.K., Fok S.-C., Peh Y.C., Ng F.C., and Sim L.S.J. Computerized detection of breast cancer with artificial intelligence and thermograms. *Journal of Medical Engineering Technology*, 26:152–157, 2002.

[45] Ng E.Y.K. and Sudharsan N. M. Effect of blood flow, tumour and cold stress in a female breast: a novel time-accurate computer simulations. *Journal of Engineering in Medicine*, 215:393–404, 2001.

[46] Ng E.Y.K. and Sudharsan N. M. An improved three-dimensional direct numerical modelling and thermal analysis of a female breast with tumour. *Journal of Engineering in Medicine*, 215:25–37, 2001.

[47] Ng E.Y.K. and Sudharsan N. M. Numerical computation as a tool to aid thermographic interpretation. *Journal of Medical Engineering & Technology*, 25:53–60, 2001.

[48] Ng E.Y.K. and Sudharsan N. M. Parametric optimization for tumour identification: bioheat equation using anova and the taguchi method. *Journal of Engineering in Medicine*, 214:505–512, 2000.

[49] Wang J., Chang K.J., Chen C.Y., Chien K.L., Tsai Y.S., Wu Y.M., Teng Y.C., and Shih T.T. Evaluation of the diagnostic performance of infrared imaging of the breast: a preliminary study. *BioMedical Engineering OnLine*, 9:1–14, 2010.

[50] Inc FLIR System. Thermovision a40m researcher. Consultado el 01/07/2012 de `http://www.movimed.com/files/Flir/A40M%20Researcher%20Datasheet.pdf`, 2005.

[51] Inc FLIR System. Thermacam researcher professional 2.9. Consultado el 01/07/2012 de `http://support.flir.com/DocDownload/Assets/47/English/T559009\protect\T1\textdollara387.pdf`, 2009.

[52] Fernández-Fernández S., Cordero-Sánchez J.M., and Córdoba-Largo A. *Estadística Descriptiva*. ESIC, 2002.

[53] Wayne W. D. *Bioestadística Base para el análisis de las ciencias de la salud*. Richard Stratton, 2009.

[54] Vargas Sabadías A. *Estadística Descriptiva e Inferencial.* Universidad de Castilla-La Mancha, 1995.

[55] Alea-Riera V., Guillén M., Muñoz C., Torrelles E., and Viladomiu N. *Estadística descriptiva: Aplicaciones prácticas.* Universidad de Barcelona, 2001.

[56] Martínez-Trinidad J. F. and Guzmán-Arenas A. The logical combinatorial approach to pattern recognition, an overview through selected works.

[57] Beyer K., Goldstein J., Ramakrishnan R., and Shaft U. When is nearest neighbor meaningful? In Catriel Beeri and Peter Buneman, editors, *Database Theory -ICDT'99*, volume 1540 of *Lecture Notes in Computer Science*, pages 217–235. Springer Berlin / Heidelberg, 1999.

[58] Weinberger K. Q., Blitzer J., and Saul L. K. Distance metric learning for large margin nearest neighbor classification. In *In NIPS.* MIT Press, 2006.

[59] Cunningham P. and Delany S.J. k-nearest neighbour classifiers. Technical report, Dublin Institute of Technology.

[60] Quinlan J. R. Induction of decision trees. In Jude W. Shavlik and Thomas G. Dietterich, editors, *Readings in Machine Learning.* Morgan Kaufmann, 1990.

[61] García-Domínguez L., Hong-Enríquez R., Sautié-Castellanos M., and J. L. Hernández-Cáceres. Complejidad y codificación en rnas no codificadores. *Revista cubana de informática médica*, 2(2), 2002.

[62] Schneider T. Information theory primer - with an appendix on logarithms, 2004.

[63] Gappmair W. Claude e. shannon: the 50th anniversary of information theory. *IEEE Communications Society*, 37:102–105, 1999.

[64] Quinlan J. R. Improved use of continuous attributes in c4.5. *CoRR*, cs.AI/9603103, 1996.

[65] Ruggieri S. Efficient c4.5. *IEEE Transactions on Knowledge and Data Engineering*, 14:438–444, 2002.

[66] Quinlan J. R. *C 4.5 Programs for Machine Learning.* Morgan Kaufmann, 1992.

[67] Freund Y. and Schapire R.E. Experiments with a New Boosting Algorithm. In *International Conference on Machine Learning*, pages 148–156, 1996.

[68] Freund Y. and Schapire R.E. A decision-theoretic generalization of on-line learning and an application to boosting. In *Proceedings of the Second European Conference on Computational Learning Theory*, EuroCOLT '95, pages 23–37, London, UK, UK, 1995. Springer-Verlag.

[69] Freund Y. and Schapire R. A short introduction to boosting. *J. Japan. Soc. for Artif. Intel.*, 14(5):771–780, 1999.

[70] Gómez-Verdejo V. *Nuevos criterios de ayuda para conjuntos de decisores cooperativos*. PhD thesis, Universidad Carlos III de Madrid. Departamento de Teoría de la Señal y Comunicaciones, May 2007.

[71] Zhu J., Rosset S., Zou H., and T. Hastie. Multi-class adaboost. Technical report, Stanford University, 2005.

[72] Pearl J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann series in representation and reasoning. Morgan Kaufmann Publishers, 1988.

[73] Neuberg L.G. Causality: Models, reasoning, and inference, by judea pearl, cambridge university press, 2000. *Econometric Theory*, 19:675–685, 2003.

[74] Friedman N. and Goldszmidt M. Learning bayesian networks from data. *University of California, Berkeley and Stanford Research Institute*, page 117, 1998.

[75] Cooper G. An overview of the representation and discovery of causal relationships using bayesian networks. *Computation Causation Discovery*, page 362.

[76] Duda R.O., Hart P.E., and Stork D.G. *Pattern Classification*. John Wiley & Sons, 2nd edition, 2001.

[77] Friedman N., Geiger D., and Goldszmidt M. Bayesian network classifiers. *Machine Learning*, 29:131–163, 1997.

[78] Russell S.J. and Norvig P. *Artificial Intelligence: A Modern Approach*. Prentice Hall, 3rd edition, 2009.

[79] Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. pages 1137–1143. Morgan Kaufmann, 1995.

[80] Han J. and Kamber M. *Data Mining: Concepts and Techniques*. The Morgan Kaufmann Series in Data Management Systems. Elsevier, 2006.

[81] Lavrac N. Selected techniques for data mining in medicine. *Artificial Intelligence in Medicine*, 16:3–23, 1999.

[82] Cross S.S., Dubé A.K., Johnson J.S., McCulloch T.A., Quincey C., Harrison R.F., and Ma Z. Evaluation of a statistically derived decision tree for the cytodiagnosis of fine needle aspirates of the breast (fnab). *Cytopathology*, 9:178–187, 1998.

[83] Cross S.S., Stephenson T.J., and Harrisont R.F. Validation of a decision support system for the cytodiagnosis of fine needle aspirates of the breast using a prospectively collected dataset from multiple observers in a working clinical environment. *Cytopathology*, 11:503–512, 2000.

[84] Cross S.S., Downs J., Drezet P., Ma Z., and Harrison R.F. *Which decision support technologies are appropriate for the cytodiagnosis of breast cancer?*, pages 265–295. World Scientific, 2000.

[85] Carver R. H. and Nash J. G. *Doing Data Analysis with SPSS, Version 18: Version 18.0*. Limusa Wiley, 2011.

[86] Sardanelli F., Giuseppetti G.M., Panizza P., Bazzocchi M., Fausto A., Simonetti G., Vincenzo Lattanzio, and Del-Maschio A. Sensitivity of mri versus mammography for detecting foci of multifocal, multicentric breast cancer in fatty and dense breasts using the whole-breast pathologic examination as a gold standard. *AJR Am J Roentgenol*, 183(4):1149–57, 2004.

[87] García E. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering In Knowledge and Data Engineering*, 21(9):1263–1284, 2009.