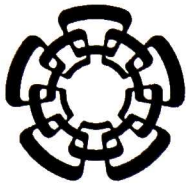
UT-TOYOTA-001
Dec. - 2015



CENTRO DE INVESTIGACIÓN Y DE ESTUDIOS AVANZADOS
DEL INSTITUTO POLITÉCNICO NACIONAL

Laboratorio de Tecnologías de Información,
CINVESTAV-Tamaulipas

Estudio comparativo de técnicas de selección de características para la clasificación de lesiones de mama en ultrasonografía

Tesis que presenta:

Cristhian Muñoz Meza

Para obtener el grado de:

Maestro en Ciencias en Computación

Director de la Tesis:
Dr. Wilfrido Gómez Flores

Cd. Victoria, Tamaulipas, México. Febrero, 2014

La tesis presentada por Cristhian Muñoz Meza fue aprobada por:

Dr. Iván López Arévalo

Dr. César Torres Huitzil

Dr. Wilfrido Gómez Flores, Director

Cd. Victoria, Tamaulipas, México., 7 de Febrero de 2014

A mis padres

Agradecimientos

- Quiero agradecer primeramente a Dios y a la vida por permitirme llegar hasta aquí.
- A mis padres por el gran amor y apoyo incondicional que siempre me han brindado.
- A mis hermanos Jorge Iván y Luis Angel por sus palabras de aliento.
- A mi hermana y su esposo por todas las facilidades prestadas, pero sobre todo por el apoyo moral en ciertas etapas de esta aventura.
- Al Dr. Wilfrido Gómez Flores por todo el apoyo y asesoría brindada, además de su paciencia.
- A mis revisores, el Dr. César Torres Huitzil y el Dr. Iván López Arévalo por sus observaciones y recomendaciones.
- A mis compañeros por su amistad y apoyo.
- Al personal administrativo por su disponibilidad y servicio eficiente brindado, y de manera muy especial a la Lic. Veronica Andrea Nava García.
- Al Consejo Nacional de Ciencia y Tecnología (CONACyT) por el apoyo financiero ofrecido.
- Al CINVESTAV por permitirme ser parte de esta comunidad.

Índice General

Índice de Figuras

Índice de Tablas

Índice de Algoritmos

Publicaciones

Cristhian Muñoz Meza and Wilfrido Gómez Flores. *A Feature Selection Methodology for Breast Ultrasound Classification* , in 10th International Conference on Electrical Engineering Computing Science and Automatic Control (CCE 2013), IEEE México City, México, September-October, 2013.

Resumen

Estudio comparativo de técnicas de selección de características para la clasificación de lesiones de mama en ultrasonografía

por

Cristhian Muñoz Meza

Laboratorio de Tecnologías de Información, CINVESTAV-Tamaulipas
Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional, 2014
Dr. Wilfrido Gómez Flores, Director

En este trabajo se propone una metodología para la selección de características en la clasificación de lesiones de mama en ultrasonografía (USM) basado en un estudio comparativo entre las técnicas de análisis de componentes principales (PCA) e información mutua (MI). Para ello se implementaron diversas técnicas de descripción morfológica y de textura propuestas en la literatura especializada, a partir de las cuales se construyeron tres espacios de características M-dimensionales: morfología (M=22), textura (M=502) y una combinación de ambos (M=524). Cada uno de estos conjuntos fue normalizado en el rango [-1,1] y, posteriormente, las características fueron ordenadas de acuerdo a su relevancia mediante las técnicas de PCA y MI. Para cada espacio M-dimensional se determinó el mejor subconjunto de características mediante la minimización del error *bootstrap* .632+ en un proceso de clasificación incremental, es decir, agregando una a una cada característica ordenada hasta que se haya considerado el conjunto completo. Una vez determinado el conjunto reducido de características para cada espacio probado, se evaluó el desempeño de cada uno de ellos mediante la métrica de área bajo la curva ROC (Az). Posteriormente, se realizó la prueba de Shapiro-Wilk ($\alpha = 0.05$) para determinar la normalidad de los datos, donde se observó que algunos grupos presentaron distribución asimétrica, por lo que se optó utilizar estadísticos robustos como la mediana (MD) y el estimador Qn.

Finalmente, se comparó la capacidad de discriminación del conjunto de características completo, con M atributos, y los subconjuntos determinados por PCA y MI, con m atributos (donde $m < M$) mediante la prueba estadística de Kruskal-Wallis ($\alpha = 0.05$).

Abstract

Comparative study of feature selection techniques for breast ultrasound classification

by

Cristhian Muñoz Meza

Information Technology Laboratory, CINVESTAV-Tamaulipas

Center for Research and Advanced Studies from the National Polytechnic Institute, 2014

Dr. Wilfrido Gómez Flores, Advisor

In this work we propose a feature selection methodology for breast ultrasound classification based in a comparative study between Principal Component Analysis (PCA) and Mutual Information (MI) techniques. There were implemented several morphological and texture descriptors proposed in literature, from which three M-dimensional features spaces were constructed: morphological ($M = 22$), texture ($M = 502$) and combined ($M = 524$). Each of these sets was normalized in the range [-1,1] and the features were ranked according their relevance through the techniques of PCA and MI. For each M-dimensional space, the best feature subset is determined by minimizing the .632+ bootstrap error by means of an incremental classification process, that is, adding one by one each ordered feature until all of them were considered. Once the reduced feature set was determined for each tested space, the performance was assessed by the metric of area under the ROC curve (Az). Subsequently, the Shapiro-Wilk test ($\alpha = 0.05$) was performed to determine the normality of the data, where it was observed that some groups showed asymmetrical distribution, so we chose to use robust statistics such as the median value (MD) and Qn estimator.

Finally, it was compared the discrimination power of the complete feature set (with M features) and the reduced spaces by PCA and MI approaches (with m features) where ($m < M$) by using Kruskal-Wallis test ($\alpha = 0.05$). Concerning texture features the Az median value was improved by increasing from 0.588, for the complete set ($M = 502$), to 0.840, for PCA ($m = 65$), and 0.820, for MI ($m = 24$). Similarly, the performance of the combined set was improved by increasing Az=0.657,

for the complete set (M = 524), to Az =0.941, for PCA (m = 69), and Az=0.951, for MI (m = 13). On the other hand, regarding the morphological features the Az median remained almost invariant, since the performance results were 0.948, for the complete set (M=22), 0.946, for PCA (m=13), and 0.943, for MI (m=3).

Nomenclatura

PCA	Análisis de componentes principales
PC	Componente principal
MI	Información mutua
mRMR	Mínima-redundancia-máxima-relevancia
US	Ultrasonografía
USM	Ultrasonografía de mama
IARC	Agencia Internacional de Investigación en Cáncer
OMS	Organización Mundial de la Salud
CAD	Diagnóstico asistido por computadora
LDA	Análisis lineal discriminante
SVM	Máquina de soporte vectorial
ANN	Redes neuronales artificiales
NRL	Longitud radial normalizada
EE	Elipse equivalente
CC	Curva de complejidad
ROI	Región de interés
GLCM	Matriz de co-ocurrencia de los niveles de gris
ROC	Característica operativa del receptor
VP	Verdadero positivo
VN	Verdadero negativo
FP	Falso positivo
FN	Falso negativo

1

Introducción

El cáncer es una enfermedad que se caracteriza por la multiplicación rápida de células que se extienden más allá de sus límites habituales y puede invadir partes adyacentes del cuerpo o propagarse a otros órganos (fenómeno conocido como metástasis). Así pues, el cáncer de mama se forma en los tejidos de la glándula mamaria, por lo general en los conductos galactóforos y los lobulillos (glándulas productoras de leche). Esta patología puede presentarse tanto en hombres como en mujeres, aunque el cáncer de mama masculino no es frecuente.

Actualmente, el cáncer de mama se ha convertido en la principal causa de muerte a nivel mundial, afectando al 16 % de la población femenina que padeció algún tipo de neoplasia maligna. Las últimas estadísticas reportadas por la Agencia Internacional de Investigación del Cáncer (IARC, por sus siglas en inglés), revelaron que en el año 2008 aparecieron alrededor de 1,384,155 nuevos casos en el mundo [1].

En México, para el año 2008, la incidencia de cáncer de mama fue de 7.57 casos por cada 100 mil habitantes, afectando principalmente a las mujeres, quienes presentan una incidencia de 14.63[1] frente a 0.27[1] de los varones. El Instituto Nacional de Estadística y Geografía (INEGI) reveló que en nuestro país la mayor incidencia se presenta en el Distrito Federal (45.84[1]), seguida por Sinaloa (45.76[1]) y San Luis Potosí (45.20[1]) [2].

Debido a que las causas del cáncer de mama aún se mantienen desconocidas, la clave para reducir la tasa de morbilidad es la detección oportuna. Es por eso que la Organización Mundial de la Salud (OMS) recomienda a los gobiernos de cada país incluir estrategias como el diagnóstico temprano y el tamizaje en sus programas de salud para la detección del cáncer de mama.

En México se emplean actividades de educación a la población y al personal de instituciones de salud a fin de identificar los síntomas en etapas tempranas de la enfermedad, las cuales están enfocadas principalmente a la difusión de la autoexploración mamaria y a la realización de estudios mamográficos periódicos.

Existen varias técnicas para la detección del cáncer de mama, dentro de las cuales destacan las siguientes:

- **Autoexploración.** Esta técnica es capaz de detectar lesiones mamarias mayores a 1 cm, basándose en la palpación y observación que realiza la mujer sobre sus propias mamas. Se recomienda practicarla mensualmente una vez que ha aparecido la menarca [3]. Su principal desventaja es que puede detectar la lesión en un estado avanzado de crecimiento.

- **Mamografía.** Es la técnica de detección por imagenología más efectiva en etapas tempranas del cáncer [4]. Se basa en la obtención de una imagen plana de la glándula mamaria mediante

[1]Número de casos por cada 100 mil habitantes.

rayos-X. La mamografía es capaz de detectar lesiones no palpables menores a 0.5 cm. Sin embargo, tiene algunas limitaciones como la realización de biopsias innecesarias debido a su baja especificidad, lo que provoca un incremento en los costos y además somete a los pacientes a una presión emocional.

- **Ultrasonido de mama.** Es la técnica coadyuvante más importante a la mamografía para la detección de lesiones de mama [4], debido a que puede visualizar la estructura interna del tejido mamario. Esta técnica consiste en la generación de imágenes basándose en las diferentes intensidades de retorno producidas por las ondas acústicas de alta frecuencia (> 7.5MHz) que se emiten sobre el tejido. Se ha demostrado que el ultrasonido de mama (USM) es capaz de distinguir entre lesiones de mama benignas y malignas basado en la textura y morfología que dichas lesiones presentan [5].

Una de las principales desventajas del ultrasonido de mama es que presenta mayor dependencia al operador que la mamografía. Se requieren de radiólogos expertos adecuadamente entrenados para poder adquirir e interpretar las imágenes de USM. Adicionalmente, en el diagnóstico de USM existen factores que afectan el desempeño del radiólogo, ya que las características sonográficas que diferencian una lesión benigna de un carcinoma pueden estar traslapadas, lo que origina las siguientes discrepancias:

- **Variación interobservador.** Se refiere a la diferencia de opiniones entre distintos radiólogos acerca de una misma imagen debido principalmente a la experiencia y entrenamiento que posean [6].

- **Variación intraobservador.** Se refiere a que un mismo radiólogo puede emitir opiniones diferentes sobre una misma imagen en momentos distintos, debido a diversos factores, como estrés, cansancio, etc. [7].

Debido a las variaciones inter/intraobservador descritas anteriormente, se han propuesto sistemas de diagnóstico asistido por computadora (CAD, por sus siglas en inglés), los cuales son desarrollados

para reducir, o inclusive eliminar, la subjetividad humana. El objetivo principal de dichos sistemas es analizar las imágenes médicas mediante algoritmos, a fin de ayudar al radiólogo en su interpretación, aumentando la sensibilidad[1] y especificidad[2]. Aunque actualmente se han presentado diversos enfoques para el desarrollo de sistemas CAD, todos ellos se basan por lo general en el esquema de la Figura 1.1 [8], que incluyen las siguientes etapas funcionales:

- **Preprocesamiento.** Una de las principales limitaciones del USM es el bajo contraste, así como la contaminación con el artefacto *speckle*, que es una propiedad inherente de las imágenes de ultrasonido que se modela como ruido multiplicativo [9]. La tarea del preprocesamiento consiste en mejorar el contraste y disminuir el *speckle* sin distorsionar las características importantes de la imagen.

- **Segmentación.** Esta fase consiste en dividir la imagen en dos regiones no traslapadas correspondientes a la lesión y el fondo.

- **Extracción y selección de características.** Una vez segmentada la lesión, se deberán extraer características que describan su morfología (forma de la lesión) y su textura (variación entre niveles de gris). Sin embargo, algunos de estos atributos pueden llegar ser irrelevantes o redundantes, de modo que es recomendable seleccionar aquellas características que aporten mayor información discriminante con respecto al tipo de lesión: benigna o maligna (también denominada como carcinoma).

- **Clasificación.** Basándose en las características seleccionadas se deben clasificar los tumores sospechosos como malignos o benignos. Para esto existen diferentes técnicas de clasificación, como son las redes neuronales artificiales (ANN), máquinas de soporte vectorial (SVM), análisis lineal discriminante (LDA), por mencionar algunas.

[1] Probabilidad de clasificar correctamente a un individuo enfermo.
[2] Probabilidad de clasificar correctamente a un individuo sano.

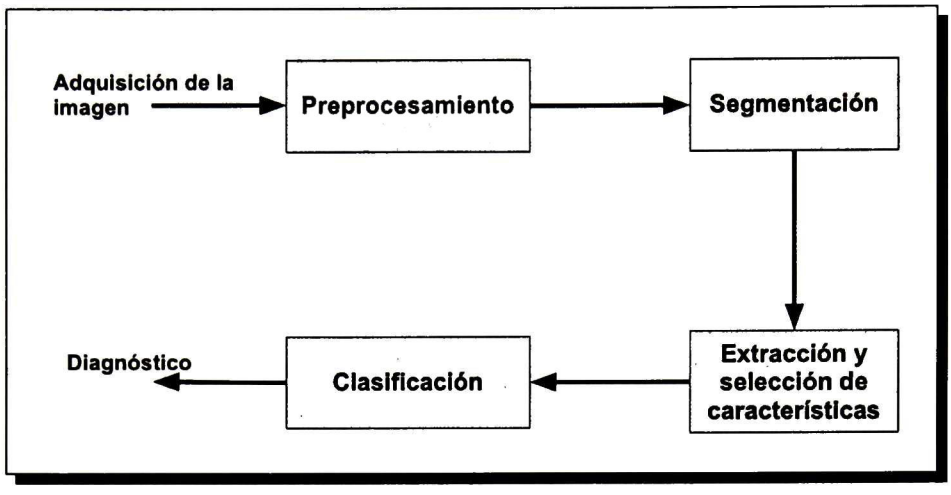


Figura 1.1: Etapas funcionales de un sistema CAD.

1.1 Planteamiento del problema

Las características sonográficas de una lesión de mama (previamente segmentada) pueden representarse numéricamente mediante atributos morfológicos y de textura. Sin embargo, el espacio de características generado suele ser de alta dimensionalidad (mayor a 20 [10]), en donde existen atributos que pueden ser irrelevantes y/o redundantes, los cuales pueden reducir el desempeño de clasificación.

Formalmente el problema de selección de características se define como:

Definición 1 (Problema general).

Dado un conjunto de datos de entrada en el espacio $\mathbb{R}$ organizado en una matriz de N muestras por M características, $X = \{x_i, i = 1, ..., M\}$, y una variable de clase (o salida deseada) c para cada muestra. El problema de selección de características es encontrar a partir de un espacio M-dimensional, $\mathbb{R}^M$, un subespacio de m características, $\mathbb{R}^m$, que caracterice adecuadamente a c, donde $m < M$.

Entonces, en esta tesis se plantea el siguiente problema de selección de características en sistemas CAD para USM:

Definición 2 (Problema planteado).

Dado un conjunto de N ultrasonografías de lesiones de mama, donde cada muestra del conjunto se representa por medio de un vector de M características, las cuales pueden ser morfológicas, textura o una combinación de ambas; y la variable de clase c puede tomar alguna de las dos categorías de lesión: benigno o maligno. Se desea encontrar el subconjunto de m características más relevantes y menos redundantes, donde $m < M$, mediante una técnica de selección de características de manera que mejore el desempeño de clasificación de lesiones en benigno y maligno.

1.2 Hipótesis

Hipótesis 1.

Dado un banco de imágenes de USM previamente segmentado, existe un subconjunto de descriptores numéricos, morfológicos y textura o una combinación de ambos, obtenido a partir de un método de selección de características que reduce el error de clasificación entre las clases lesión benigna y carcinoma.

1.3 Objetivos generales y específicos

1.3.1 General

Definir una metodología para obtener el subconjunto de atributos morfológicos y de textura que describan lesiones de mama en ultrasonografías, a partir de un estudio comparativo de técnicas de

selección de características, de manera que se reduzca el error de clasificación.

1.3.2 Particulares

1. Realizar un estudio de las diferentes técnicas de descripción (morfología y textura) propuestas en la literatura para la clasificación de lesiones de mama en ultrasonografías.

2. Construir tres espacios de características M-dimensionales (morfológicas, textura y combinación de ambos) a partir de un conjunto de ultrasonografías de mama previamente segmentadas.

3. Determinar el subconjunto de características que proporcione mayor información discriminante entre las clases de lesión benigna y maligna, mediante la comparación de distintas técnicas de selección de características.

4. Definir un esquema de validación de los subconjuntos de atributos obtenidos en términos del desempeño de un clasificador.

1.4 Metodología

La presente investigación está dividida en cuatro etapas fundamentales:

1. **Segmentación.** Se dividirá la imagen de USM en dos regiones disjuntas, lesión y fondo, mediante una técnica basada en la transformada *watershed* [11].

2. **Extracción y evaluación de características.** Se implementarán diversos descriptores morfológicos y de textura propuestos en la literatura, con el fin de determinar aquellos que aportan mayor información discriminante entre las clases de lesión benigna y maligna.

3. **Selección de características.** Se implementarán diferentes técnicas de selección de características, comúnmente utilizadas en CADs para ultrasonografía de mama, con el fin de

determinar aquella que proporcione un subespacio de características donde la dimensionalidad del espacio original se reduzca y que al mismo tiempo produzca el menor error de clasificación.

4. **Evaluación de resultados.** Se implementará un clasificador basado en el análisis lineal discriminante de Fisher (FLDA, por sus siglas en inglés), con el fin de evaluar el desempeño de clasificación del subconjunto de atributos (morfológicos, textura o combinado) elegido por cada una de las técnicas de selección de características implementadas previamente. Dicha evaluación se llevará a cabo mediante el análisis de la característica operativa del receptor (ROC, por sus siglas en inglés) mediante la métrica de área bajo la curva (Az).

En la Figura 1.2 se ilustra el esquema global de la metodología que se seguirá en el desarrollo de esta investigación.

1.5 Organización del trabajo de tesis

El presente trabajo de tesis está dividido en seis capítulos. En el Capítulo 1 se da una introducción a fin de contextualizar la importancia de la investigación sobre el cáncer de mama, además se describen los principales métodos de diagnóstico para este padecimiento, ofreciendo así una justificación para el desarrollo de este trabajo. En el Capítulo 2 se describen los principales trabajos relacionados con el trabajo de tesis. En el Capítulo 3 se describen de manera teórica los diferentes métodos y técnicas utilizados en el desarrollo de este trabajo. En el Capítulo 4 se describe la metodología propuesta para la comparación de dos técnicas de selección de características para la clasificación de lesiones de mama en USM. En el Capítulo 5 se presentan los resultados obtenidos con el método propuesto. Por último en el Capítulo 6 se presentan las conclusiones así como el trabajo futuro.

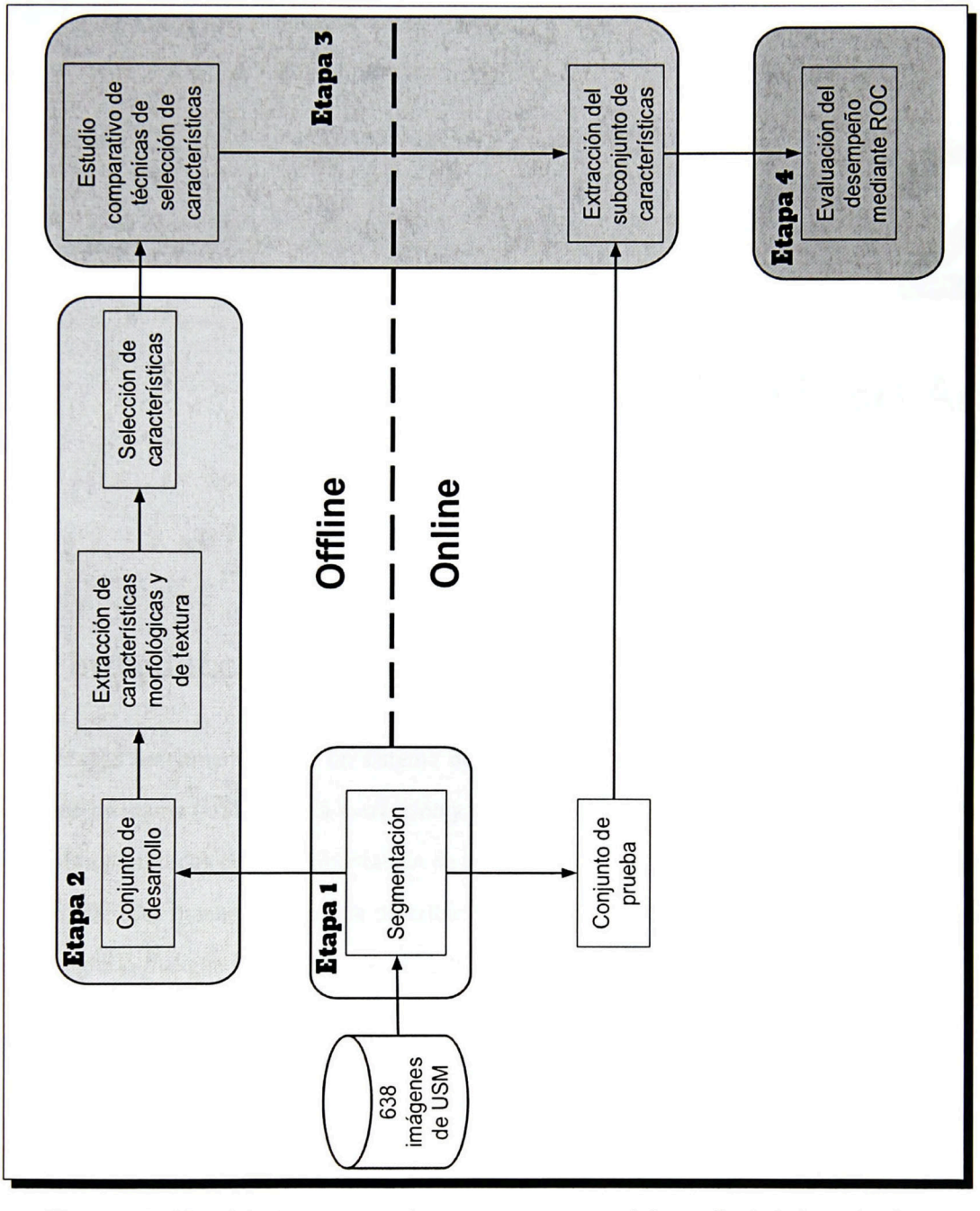


Figura 1.2: Metodología propuesta de cuatro etapas para el desarrollo de la investigación.

2

Estado del Arte

2.1 Introducción

Dos etapas fundamentales en un sistema de diagnóstico asistido por computadora (CAD) para ultrasonido de mama (USM) son la extracción y la selección de características de los tumores. Ambas etapas están precedidas por la segmentación de la lesión, donde la región del tejido tumoral es aislada del fondo. De esta manera, es posible describir la naturaleza de la lesión de interés para clasificarla como benigna o maligna [12].

La extracción de características consiste en cuantificar atributos (también conocidos como descriptores, rasgos o parámetros) que ayuden a diferenciar numéricamente a una lesión benigna de un carcinoma. Para este propósito, en la literatura se han propuesto una gran variedad de descriptores de textura y morfológicos. Una textura se interpreta como la variación del patrón de intensidad en escalas menores a las escalas de interés. En un sistema CAD para USM la idea básica es medir la

heterogeneidad de los niveles de gris dentro y fuera de la región tumoral. Por otra parte, la descripción de la morfología de una lesión ayuda a cuantificar el grado de irregularidad de su contorno y además se pueden calcular otros parámetros como orientación, relación de aspecto, relación de áreas, etc.

La selección de características determina, a partir de un conjunto completo de descriptores, un subconjunto de rasgos de textura y/o morfológicos que incrementen el desempeño de clasificación. De esta manera se reduce la dimensionalidad del espacio de características, lo cual impacta en el tiempo de cómputo de la clasificación, y se establece un subconjunto de características relevantes, es decir, que aportan la mayor distinción numérica entre las dos clases: **lesión benigna** y **carcinoma**.

Diversos autores han contribuido con varias técnicas para extraer rasgos de textura y morfológicos relevantes, para la clasificación de lesiones de mama. A continuación se describen los trabajos más relevantes y actuales para resolver las etapas de extracción y selección de características.

2.2 Descriptores de textura

Chang *et al.* [13] propusieron un método para la clasificación de lesiones de mama que combina los coeficientes de auto-covarianza con la información del *speckle*. Primeramente implementaron un detector para localizar píxeles de *speckle* y posteriormente se calculan los coeficientes de auto-covarianza para dichos píxeles. Cada imagen de ultrasonido produjo una matriz de auto-covarianza del *speckle* de tamaño 5x5. Sin embargo, debido a que el coeficiente (0,0) siempre es la unidad para la matriz de auto-covarianza normalizada, éste es descartado, obteniendo 24 coeficientes, formando así un vector de características 24-dimensional. El conjunto total de imágenes se dividió en cinco grupos, utilizando el primer grupo como conjunto de prueba y los otros cuatro para entrenamiento de un clasificador basado en SVM. Una vez entrenado el clasificador, éste es probado con el primer grupo. El experimento se repitió hasta utilizar los cinco grupos como conjunto de prueba.

Parámetros calculados a partir de la curva de complejidad (CC) han demostrado su capacidad para distinguir entre tumores de mama con texturas homogéneas o heterogéneas. Alvarenga *et al.* [14] emplearon la CC y la matriz de co-ocurrencia de niveles de gris (GLCM, por sus siglas en inglés). Se definieron dos regiones de interés (ROI, por sus siglas en inglés), donde la primera incluía un área que contenía la lesión y parte del fondo, mientras que la segunda únicamente consideraba la región interna de la lesión. Para ambas ROI se computaron cinco parámetros (valor máximo de transiciones, valor medio de transiciones, media de la muestra, desviación estándar de la muestra y entropía) extraídos de la CC y cinco parámetros (contraste, entropía, desviación estándar, segundo momento angular y correlación) obtenidos de la GLCM. De esta manera se obtuvo un vector 20-dimensional de características de textura. Se utilizó el análisis lineal discriminante de Fisher (FLDA, por sus siglas en inglés) como clasificador para evaluar el desempeño de clasificación de los descriptores de manera individual así como para combinaciones de hasta cinco de ellos.

Del mismo modo que en el trabajo de Alvarenga *et al* [14], Liao *et al.* [15] hicieron uso de la GLCM para describir la textura de la lesión en imágenes de ultrasonografía de mama (USM). Once radiólogos expertos delinearon el contorno de la lesión de manera manual. Se calcularon cuatro GLCMs con diferentes direcciones ($\theta = 0°$, $45°$, $90°$ y $135°$) a fin de evitar el posible sesgo direccional. Para cada GLCM se calcularon cuatro parámetros (homogeneidad, contraste, energía y varianza), haciendo un total de 16 descriptores de textura para cada imagen. Se usó el FLDA como clasificador y se evaluó el desempeño mediante el análisis ROC.

Liu *et al.* [16] propusieron un método completamente automático para la clasificación de lesiones de mama en ultrasonografía. Este método fue dividido en dos fases principales: 1) generación automática de la ROI y 2) clasificación de la ROI. En la primera fase se divide la imagen utilizando una retícula donde los elementos son cuadrados del mismo tamaño y se extraen las características de

textura basadas en la GLCM. A continuación se clasifica cada elemento de la retícula como región de tejido normal o región candidata de lesión de mama mediante una SVM con función núcleo de base radial.

En la segunda fase se distribuyen algunos puntos de clasificación en cada ROI. Para cada punto se forman cinco ventanas alrededor y se calcula la GLCM de cada una considerando cinco distancias ($d = 1, 2, 3, 4$ y 5) y cuatro orientaciones ($\theta = 0°$, $45°$, $90°$ y $135°$). Posteriormente se obtienen cuatro descriptores (entropía, contraste suma de promedios y suma de entropía). Cada punto es clasificado mediante SVM y se obtiene la razón entre la cantidad de puntos malignos y el total de puntos, la cual es comparada con un umbral predefinido para determinar si la lesión es benigna, o se trata de cáncer.

2.3 Descriptores morfológicos

R-F Chang *et al.* [17] utilizaron seis descriptores morfológicos (factor de forma, redondez, relación de aspecto, convexidad, solidez y extensión) para la caracterización de lesiones de mama. Debido que el artefacto *speckle* degrada la calidad de las imágenes de ultrasonido, los autores primeramente mejoraron el contraste de la imagen así como los bordes de la lesión mediante la aplicación de un filtro de difusión anisotrópico y el método *stick*[18]. Para la segmentación se implementó el método de umbralado automático propuesto por Otsu [19]. Además se incluyó un sistema de control donde el usuario pudiera cambiar el umbral cuando no esté de acuerdo con el umbral asignado automáticamente. Una SVM con *kernel* Gaussiano de base radial fue utilizado como clasificador.

El contorno de la lesión proporciona información relevante para la clasificación de lesiones de mama. Huang *et al.* [20] utilizaron 19 descriptores morfológicos extraídos del borde de la lesión, los cuales se muestran en la Tabla 2.1 . Los autores implementaron un método de segmentación para la extracción automática del contorno de la lesión en imágenes de ultrasonografía [21]. Dicho método

Número	Descriptor
1	Perímetro
2	Área
3	Número de protuberancias y lobulaciones sustanciales
4	Índice de lobulación
5	Circunferencia elíptico-normalizada
6	Esqueleto elíptico normalizado
7	Relación eje mayor - eje menor
8	Relación de aspecto
9	Factor de forma
10	Redondez
11	Solidez
12	Convexidad
13	Extensión
14	Relación de áreas tumor-envolvente convexa
15	Relación de perímetros tumor-elipse
16	Diferencia de perímetros tumor-elipse
17	Relación de perímetros tumor-circunferencia
18	Diferencia de perímetros tumor-circunferencia
19	Relación área-perímetro

Tabla 2.1: Descriptores utilizados por Huang *et al.*

consiste en aplicar un filtro denominado ecuación de difusión de curvatura modificada (MCDE, por sus siglas en inglés) para mejorar la imagen, posteriormente se aplica un método de umbralado automático, el cual minimiza la variación inter-clase entre píxeles blancos y negros. La clasificación se realizó mediante SVM.

Una parte esencial en la clasificación de lesiones de mama es la determinación de la ROI. Alvarenga *et al.*[22] proponen un método de clasificación donde primeramente se obtiene el contorno de la lesión mediante un método semiautomático basado en operadores morfológicos y la transformada *watershed*.

Una vez delimitada la lesión, se extrajeron siete descriptores morfológicos. Los tres primeros fueron la desviación estándar, relación de área y rugosidad del contorno, los cuales fueron obtenidos

a partir de la longitud radial normalizada. Los siguientes dos descriptores fueron la relación de área y el valor residual normalizado, que fueron calculados a partir del polígono convexo. Finalmente, se calcularon dos parámetros geométricos que describen la circularidad y el cociente morfológico, los cuales son comúnmente utilizados en CADs para USM.

Gómez *et al.* [23] propusieron un sistema CAD donde calculan 22 descriptores morfológicos basados en el polígono convexo, elipse equivalente y longitud radial normalizada, los cuales cuantifican la irregularidad de los límites de la lesión. Los autores implementaron un sistema de segmentación semiautomático basado en la transformada *watershed*. Se empleó la técnica de información mutua (MI) para la selección de características mediante el criterio de máxima-relevancia-mínima-redundancia (mRMR). Se utilizaron tres clasificadores: una red neuronal con función de base radial (RBFNN, por sus siglas en inglés), SVM y FLDA.

Bocchi *et al.* [24] emplearon un método de clasificación de lesiones de mama basado en la elaboración de un video completo adquirido mediante sonografía, el cual capturaba imágenes de la lesión desde diferentes puntos de vista, a diferencia del método tradicional donde la imagen sólo muestra un plano de la lesión. Cada video es dividido en un conjunto de imágenes, las cuales son cortadas de acuerdo a la ROI, misma que se define para este contexto como el campo de visión completo del instrumento de extracción (transductor), excluyendo la información textual incluída en el video, cada uno consta de 100-150 cuadros, dependiendo del transductor. La segmentación se realizó mediante la técnica de contornos activos (*snakes*), inicializando manualmente el *snake* en el primer frame del *videoclip*. Se calcularon tres parámetros que describen las características morfológicas (forma y relación de ejes) y ecogénicas (interfaz de halo). Este método asume que cada lesión es vista de diferentes ángulos, por lo que una misma lesión podría verse como benigna o cáncer, proporcionando marcos erróneos. Para solucionar este problema se utilizó un clasificador con dos fases de entrenamiento con el objetivo de identificar los marcos erróneos. El clasificador fue

diseñado mediante una red neuronal con una capa oculta de 10 unidades, mismas que se definieron de manera experimental, donde todas las unidades fueron activadas con la función de tangente hiperbólica.

2.4 Combinación de descriptores

Wu y Moon (2008) [25] emplearon una combinación de características morfológicas y de textura. Los autores implementaron la técnica de selección de características hacia adelante (FFS, por sus siglas en inglés) para encontrar los parámetros más importantes de seis descriptores morfológicos. Este método ordenó las características de acuerdo a su importancia, encontrando que la solidez es el descriptor morfológico más relevante. Adicionalmente se calcularon los coeficientes de auto-covarianza (descriptor de textura) derivados del análisis estadístico. Previamente las imágenes fueron segmentadas, para ello primeramente se mejoró la calidad de la imagen y los bordes de esta mediante la aplicación de un filtro de difusión anisotrópico, así como el método *stick*; posteriormente se aplicó la técnica de umbralado automático propuesta por Otsu y se utilizó como clasificador SVM.

Además de los descriptores morfológicos y de textura existen otros parámetros que proporcionan información para la clasificación de lesiones mediante USM. Liao *et al.* [26] combinaron dos descriptores morfológicos (compacidad y desviación estándar de la distancia más corta), dos de textura (energía y varianza), así como un parámetro que modela la retropropagación del *speckle* basado en la distribución Nakagami. El contorno de las imágenes fue determinado de manera manual por radiólogos expertos. La clasificación se realizó mediante la técnica de c-medias difusa.

Es posible analizar el flujo de la sangre en imágenes de USM, mediante la técnica de flujo Doppler. Liu *et al.* [27] utilizaron una combinación de características morfológicas y de textura, así como características extraídas de la imagen de flujo Doppler para la clasificación de lesiones de mama.

Debido a que no todas las características son efectivas para la clasificación, se calculó la distancia de clasificación de cada descriptor entre tumor maligno y benigno a fin de evaluar la efectividad de cada parámetro de entrada. Se obtuvieron los cinco mejores conjuntos para cada tipo de descriptor. Se utilizó SVM para clasificar la lesión utilizando cada subconjunto obtenido de manera individual, así como combinaciones de ellos.

Moon *et al.* [28] utilizaron 15 descriptores morfológicos, así como 16 parámetros de textura extraídos a partir de la GLCM. Dicha información se calcula de la imagen volumétrica 3D que contiene toda la información anatómica de la lesión. Un radiólogo experto indicó la localización del tumor, posteriormente se reconstruyó dicha imagen, denominada volumen de interés (VOI, por sus siglas en inglés). Se utilizó un modelo de regresión logística binaria como clasificador, evaluando el desempeño de cada tipo de descriptor, así como una combinación de ambos.

2.5 Selección de características

Pereira *et al.* [29] evaluaron el desempeño de siete descriptores morfológicos extraídos de la longitud radial normalizada (NRL) y el polígono convexo. La NRL mide la distancia del centroide de la lesión a cada píxel del contorno, mientras que el polígono convexo se define como el área convexa más pequeña que contiene todos los puntos dentro de la lesión. La técnica de información mutua se empleó para determinar la relevancia de cada uno de los descriptores.

El RPCA es una variante de la técnica de análisis de componentes principales (PCA, por sus siglas en inglés) la cual promete mayor robustez en cuanto a observaciones atípicas o corruptas [30]. En el trabajo propuesto por Wan *et al.* [31] se considera el uso de dicha técnica. Los autores utilizan una combinación de descriptores morfológicos y de textura debido a su remarcable poder discriminativo. En este trabajo se utilizó un conjunto de imágenes donde se delineó el contorno de la lesión de manera

manual por un grupo de radiólogos expertos. Se dividió el conjunto total de imágenes en conjunto de prueba y conjunto de entrenamiento. Un total de 1285 características fueron calculadas incluyendo morfológicas y de textura. Se aplicó el RPCA para determinar el mejor conjunto de características. Posteriormente, se utilizó una SVM para la clasificación de la lesión con el conjunto seleccionado. Los autores comparan el uso de RPCA con el desempeño de PCA tradicional, demostrando que el primero obtuvo mejores resultados.

Gómez *et al.* [32] analizaron el comportamiento de 22 descriptores de textura calculados a partir de la GLCM. Primeramente se segmentó la imagen mediante un algoritmo basado en la transformada *watershed*. Posteriormente, se recortó la mínima área rectangular que contiene la lesión para extraer las características de la GLCM de esa región. Se consideraron 10 distancias ($d = 1, 2, ..., 10$) cuatro orientaciones ($\theta = 0°$, $45°$, $90°$ y $135°$) y seis cuantificaciones de niveles de gris ($L = 8, 16, 32, 64, 128, 256$), posteriormente cada GLCM es normalizada. Se utilizó la técnica de información mutua (MI, por sus siglas en inglés) con el criterio de mínima redundancia máxima relevancia (mRMR) para reducir la dimensionalidad del espacio de características y se utilizó FLDA para la clasificación de las lesiones de mama.

Un método de segmentación automático fue propuesto por Rivera y Gómez [33] donde la imagen es particionada en rejillas no traslapadas de tamaño 16x16 píxeles, calculando 16 GLCM (cuatro orientaciones con cuatro distancias y cuantificación de 64 niveles de gris), de las cuales se extraen 22 descriptores, haciendo un total de 352 características de textura. Las imágenes utilizadas fueron delineadas previamente por un experto. Se aplicaron cinco técnicas de mejoramiento de contraste y cinco de filtrado del *speckle*, generando 25 combinaciones de preprocesamiento. Para cada combinación se aplicó la técnica de información mutua con el criterio mRMR.

Wu *et al.* [34] proponen un método de clasificación automática, en donde emplearon seis

descriptores morfológicos y 24 rasgos de textura extraídos a partir de la matriz de auto-covarianza. Primeramente, procesaron la imagen con un filtro de difusión anisotrópico, después aplicaron el método *stick*, seguido de un método de umbralado automático con lo que se obtuvo un contorno inicial, el cual es deformado mediante el método de contornos activos para finalmente obtener el contorno de la lesión. Una vez segmentada la imagen se calcula la matriz de auto-covarianza de 5x5, descartando el coeficiente (0,0) debido a que su valor siempre es uno. Adicionalmente, se calcularon seis descriptores morfológicos. Se utilizó un algoritmo genético (GA) para la selección de características, y se ejecutó cinco veces reduciendo la dimensionalidad del espacio de 30 a 5, 6, 9, 10 y 8, en cada ejecución. La clasificación se llevó a cabo mediante SVM.

2.6 Técnicas de clasificación

La clasificación es la tarea que permite diferenciar entre distintas clases. En el diagnóstico de lesiones de mama, esta fase consiste en determinar si una lesión es benigna o un carcinoma. Los métodos de clasificación se pueden dividir en dos categorías: supervisados y no supervisados. En la clasificación supervisada se tiene un conocimiento previo de las clases a las que pertenecen los patrones de entrada. En la clasificación no supervisada no se tiene ningún conocimiento acerca de las clases de los patrones, sino que son agrupados de acuerdo con un criterio de similitud dado. En el desarrollo de esta investigación se utiliza clasificación supervisada debido a que todas las imágenes de USM han sido diagnosticadas mediante un estudio histopatológico. Como se observa en la Tabla 2.2, en el diagnóstico por USM es común el uso de técnicas de clasificación supervisada como el análisis lineal discriminante de Fisher (FLDA), máquina de soporte vectorial (SVM), redes neuronales artificiales (ANN) y regresión lineal (LR).

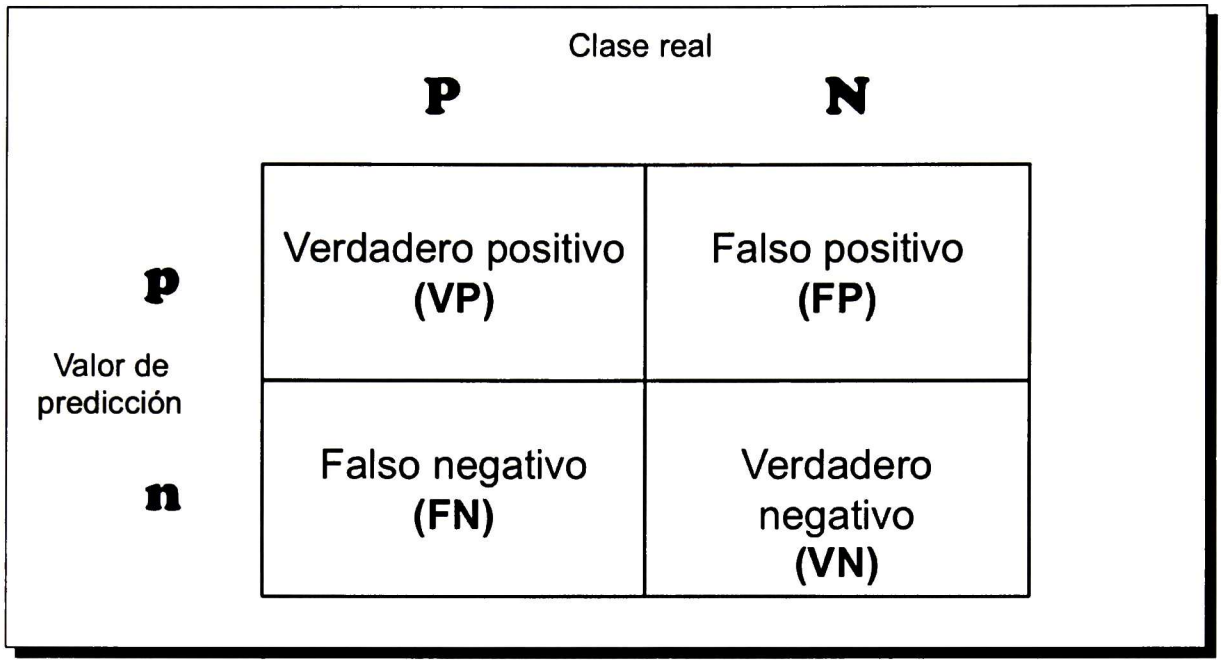


Figura 2.1: Matriz de confusión.

2.7 Análisis ROC

El análisis de las características operativas del receptor (ROC) es ampliamente usado para medir el rendimiento de sistemas de clasificación cuyas respuestas son dicotómicas [35], es decir, que una instancia clasificada puede tomar uno de dos posibles valores de clase: positivo (p) o negativo (n). En el caso de clasificación de lesiones de mama se consideran las clases benigna y maligna. Por tanto, se generan cuatro posibles relaciones entre el valor real de la instancia (denotados con mayúsculas P ó N) y el valor de predicción del clasificador (denotado con minúsculas p ó n). Estas relaciones son definidas en la matriz de confusión que se muestra en la Figura 2.1 donde:

- **VP**: verdadero positivo (diagnóstico positivo y carcinoma presente).
- **VN**: verdadero negativo (diagnóstico negativo y carcinoma ausente).
- **FP**: falso positivo (diagnóstico positivo y carcinoma ausente).

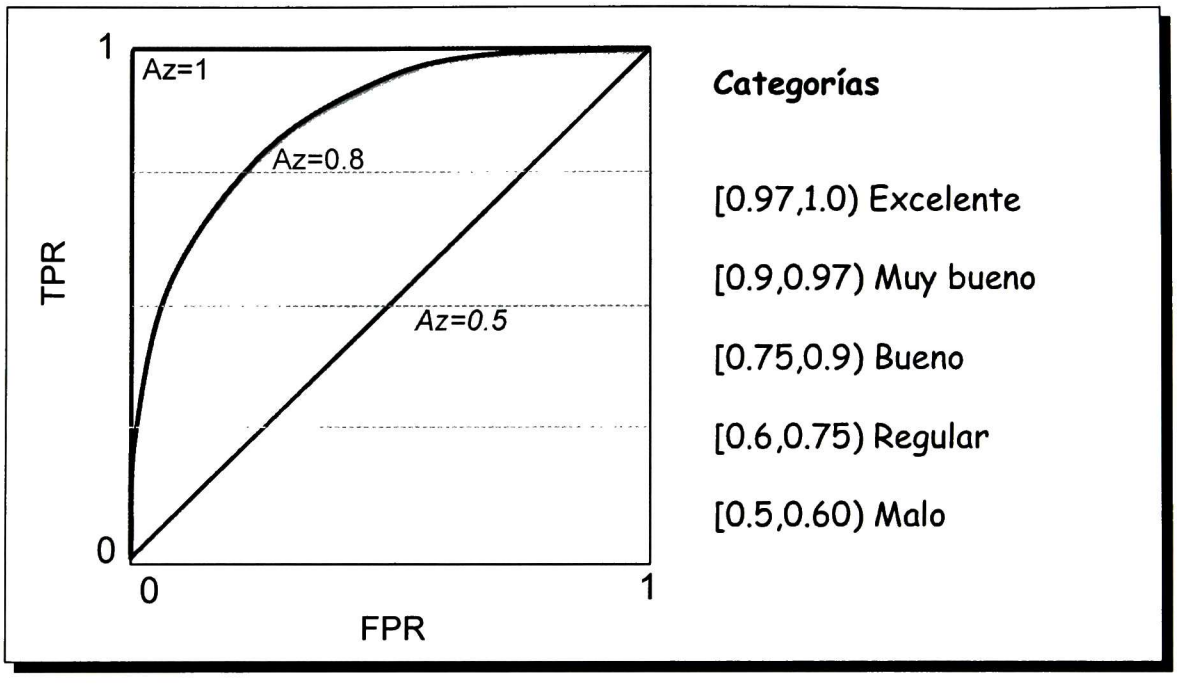


Figura 2.2: Categorías del índice de área bajo la curva ROC donde FPR indica la Razón de Falsos Positivos, mientras que TPR denota Razón de Verdaderos Positivos.

- **FN**: falso negativo (diagnóstico negativo y carcinoma presente).

El análisis ROC define ocho métricas basadas en las cuatro relaciones obtenidas de la matriz de confusión: sensibilidad (SE), especifidad (EP), probabilidad de falsa alarma (PFA), probabilidad de falsa holgura (PFH), valor de predicción positivo (VPP), valor de predicción negativo (VPN), exactitud (EX) y probabilidad de error (PE). Dichas métricas se encuentra en el rango [0,1]. Las métricas más comúnmente usadas en la medición de eficiencia de un clasificador son SE, EP, EX, así como el área bajo la curva ROC (Az), que se interpreta como la probabilidad de que un sistema de clasificación categorice a una instancia positiva más alto que a una negativa.

En el análisis ROC se definen cinco categorías que van desde un desempeño aleatorio (Az=0.5) hasta un desempeño perfecto (Az=1), como se muestra en la Figura 2.2.

En la Tabla 2.2 se hace un resumen de las características principales de los trabajos más relevantes descritos anteriormente, incluyendo sus desempeños de clasificación en términos del análisis ROC.

2.8 Conclusiones

De los artículos descritos anteriormente se concluye lo siguiente:

- La características utilizadas se categorizan en morfológicos y textura ó en algunos casos combinaciones de ellos.
- Cada autor propone su propio conjunto de características que generalmente provienen de una sóla técnica de descripción. Por tanto, existe una gran cantidad de descriptores que no se han estudiado aún en conjunto.
- Las técnicas de selección de características comúnmente utilizadas son el Análisis de Componentes Principales (PCA) y la Información Mutua (MI). Sin embargo, no se ha estudiado cuál de ambas técnicas proporciona un subconjuto de características que mejore el desempeño de clasificación.
- No existe un repositorio de imágenes estándar y cada autor propone su propio conjunto de imágenes.

Autores	REF	NI	NC	TD	MS	TC	Desempeño de clasificación (análisis ROC)
Chang *et al.*	[13]	250 (140-B, 110-M)	24	textura	—	SVM	Ac=93.20 %, Se=95.45 % y Sp=91.43 %
R-F Chang *et al.*	[17]	210 (120-B, 90-M)	6	morfológico	—	SVM	Ac=90.95 %, Se=88.89 % y Sp=92.50 %
Alvarenga *et al.*	[14]	152 (52-B, 100-M)	10	textura	—	FLDA	Ac=84 %, Se=87 % y Sp=78 %
Huang *et al.*	[20]	118 (84-B, 34-M)	19	morfológico	PCA	SVM	Ac= 82.20 %, Se=91.18 %, Sp=78.57 % y Az=0.89
Wu y Moon	[25]	210 (120-B, 90-M)	9	combinación	—	SVM	Ac=92.86 %, Se= 94.44 %, Sp=91.67 % y Az=0.94
Wang *et al.*	[36]	168 (81-B, 87-M)	18	combinación	—	ANN	Ac=94.1 %, Se=95.4 % y Sp=92.7 %
B. Liu *et al.*	[16]	112 (52-B, 60-M)	4	textura	—	SVM	Az=0.96
Alvarenga *et al.*	[29]	246 (177-B, 69-M)	7	morfológico	—	FLDA	Ac=83 %, Se=83 %, Sp=85 % y Az=0.86
Liao *et al.*	[26]	100 (50-B, 50-M)	4	combinación	—	c-means	Ac= 86 %, Se=82 % y Sp=90 %
Pereira *et al.*	[22]	246 (69-B, 177-M)	7	morfológico	MI	FLDA	Az=0.86 y Az=0.78
Liao *et al.*	[15]	100 (50-B, 50-M)	4	textura	—	FLDA	Ac= 84 %, Se= 76 %, Sp=92 % y Az=0.90
Wan *et al.*	[31]	321 (113-B, 208-M)	1285	combinación	PCA RPCA	SVM	Ac=66.8 %, Se=52.9 % y Sp=86.9 % Ac=85.9 %, Se=78.3 % y Sp=91.2 %
Gómez *et al.*	[32]	436 (219-B, 217-M)	17	textura	mRMR	FLDA	Ac=83.05 %, Se=78.02 %, Sp=88.11 % y Az=0.87
Bocchi *et al.*	[24]	30(21-B, 9-M)	3	morfológico	—	ANN	Ac=89 %, Se=89 %, Sp=90 % y Az=0.95
Liu *et al.*	[27]	105 (50-B, 55-M)	3	combinación	—	SVM	Ac=94.28 %, Se=96.36 % y Sp=92.98 %
Moon *et al.*	[28]	147 (76B-71M)		combinación	—	LR	Ac=87.8 %, Se=91.6 %, Sp=84.2 % y Az=0.96
Rivera y Gómez	[33]	960	125	textura	mRMR	SVM	Az= 0.8149, Se= 071.29 % y Sp=71.04 %
W.-J. Wu *et al.*	[34]	210 (120-B, 90-M)	5	combinación	GA	SVM	Ac=95.24 %, Se= 97.78 % y Sp=93.33 %

Tabla 2.2: Trabajos más relevantes del estado del arte. NI: Número de imágenes, donde B indica benigno y M indica Maligno, NC: Número de características, TD: Tipo de descriptor, MS: Método de selección de características, TC: Técnica de clasificación.

3

Marco teórico

3.1 Introducción

En este capítulo se describen de forma teórica las diversas técnicas empleadas en el desarrollo de esta tesis. Para ello se han preparado cuatro secciones: *extracción de características, selección de características, clasificación y estimación del error.* En la primera sección se explican las diferentes técnicas para la extracción de características que describan la lesión; en la segunda sección se definen las técnicas de selección de características implementadas para el desarrollo de este trabajo; en la tercera sección se describe la técnica de clasificación basada en el análisis lineal discriminante de Fisher (FLDA, por sus siglas en inglés); por último, se describe la técnica de estimación del error denominada *bootstrap .632+*.

3.2 Extracción de características

En un sistema CAD para USM es necesario representar las características de la lesión de manera numérica, a fin de que el clasificador pueda tomar una decisión. En esta sección se definen las diferentes técnicas de descripción utilizadas en el desarrollo de esta tesis.

3.2.1 Descriptores morfológicos

A continuación se describen diferentes técnicas que permiten evaluar de forma numérica características como tamaño, forma, orientación, etc.

3.2.1.1. Longitud radial normalizada

Sea S el área de una lesión previamente segmentada y P el perímetro de S. La longitud radial normalizada (NRL) se define como la distancia entre el centroide de la lesión hacia cada uno de los puntos en P, y se expresa como [22, 37]:

$$d_n(i) = \frac{d(i)}{\text{max}[d(i)]} \tag{3.1}$$

donde la distancia Euclidiana entre el i-ésimo píxel en P, $(x(i), y(i))$ y el centroide de S, (x_0, y_0) se define como:

$$d(i) = \sqrt{(x(i) - x_0)^2 + (y(i) - y_0)^2}, \quad 1 \leq i \leq N \tag{3.2}$$

donde N es el número de píxeles en P.

De la Ecuación 3.1 se derivan los siguientes descriptores:

1. **Desviación estándar.** Se emplea como una medida de la variación del contorno, y se expresa

como [22]:

$$D_{NRL} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(d_n(i) - \overline{d}_n)^2}, \tag{3.3}$$

donde $\overline{d}_n$ es el valor medio de $d_n(i)$.

2. **Razón de área.** Proporciona la distancia promedio entre un círculo hipotético de radio $\overline{d_n}$ y los píxeles en el borde de la lesión que se encuentran fuera de dicho contorno circular. Este parámetro se denota como [22]:

$$RA = \frac{1}{N-1} \cdot \sum_{i=1}^{N}(d_n(i) - \overline{d}_n), \tag{3.4}$$

donde $d_n(i) = 0 \forall d_n(i) \leq \overline{d}_n$.

3. **Entropía.** Representa la irregularidad del contorno de la lesión, y se expresa como [37]:

$$E = -\sum_{k=1}^{100} p_k \log(p_k) \tag{3.5}$$

donde p_k representa la probabilidad para cada percentil.

4. **Rugosidad del contorno.** Mide las diferencias entre píxeles consecutivos para la NRL, y se calcula como [22]:

$$R = \frac{1}{N}\sum_{i=1}^{N} |d_n(i) - d_n(i+1)| \tag{3.6}$$

3.2.1.2. Envolvente convexa

Se define como la región convexa más pequeña que contiene todos los puntos que pertenecen a una determinada región. Sea S_0 la envolvente convexa para una lesión S. Entre más irregular es la

lesión, más grande será la diferencia de áreas entre S y S_0. Dos parámetros son calculados a fin de cuantificar esta característica [22]:

1. **Relación de área:**

$$RS = \frac{Area(S \cap S_0)}{Area(S \cup S_0)} \tag{3.7}$$

2. **Valor residual normalizado:**

$$nrv = \frac{Area(S_0) - Area(S)}{p_0} \tag{3.8}$$

donde p_0 es el perímetro de S_0.

Sea p_i el i-ésimo punto en el contorno de la lesión, ordenados en sentido horario. Para cada punto p_i se define su profundidad h_i como la distancia más corta entre p_i y la envolvente convexa. Un punto p_i se define como cóncavo si h_i rebasa un umbral θ_1, de igual manera un punto p_i se define como convexo si h_i es menor a un umbral θ_2. Si dos puntos cóncavos no contienen un punto convexo entre ellos, entonces el punto con menor h_i es eliminado, similarmente si dos puntos convexos no contienen un punto cóncavo entre ellos, el punto con mayor h_i es eliminado.

Entonces, dados el conjunto de puntos cóncavos $\Omega = \{\omega_1, \omega_2, \ldots, \omega_d\}$ y el conjunto de puntos convexos $\Lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_p\}$, es posible determinar el **número de protuberancias y depresiones sustanciales** ($NSPD$) como [38]:

$$NSPD = p + d \tag{3.9}$$

donde d es el número de puntos cóncavos y p el número de puntos convexos.

El **índice de lobulación** (LI) es usado para describir la distribución de los lóbulos en una lesión. Como se observa en la Figura 4.5, un lóbulo se define como la región en gris delimitada por

el contorno de la lesión y la línea punteada que conecta dos puntos cóncavos adyacentes. Supongase una lesión con N_I lóbulos y el área del i-ésimo lóbulo es A_i, para $i = 1, 2, \ldots, N_I$, entonces LI se calcula como [38]:

$$LI = \frac{A_{max} - A_{min}}{\frac{1}{N_I}\sum_{i=1}^{N_I} A_i} \tag{3.10}$$

donde A_{max} y A_{min} representan las áreas de los lóbulos mayor y menor, respectivamente.

3.2.1.3. Elipse equivalente

Considérese la ecuación de la elipse con centro en el origen $(0, 0)$, definida por:

$$E = \{(x, y) : ax^2 + 2bxy + cy^2 = 1\} \tag{3.11}$$

donde a,b y c son coeficientes. Es posible encontrar una elipse de igual área y centro de masa que una lesión mediante el cálculo de sus momentos de segundo orden, definidos como:

$$S_{xx} = \frac{1}{A}\sum_{(x,y)\in A} (x - x_0)^2 \tag{3.12}$$

$$S_{yy} = \frac{1}{A}\sum_{(x,y)\in A} (y - y_0)^2 \tag{3.13}$$

$$S_{xy} = \frac{1}{A}\sum_{(x,y)\in A} (x - x_0)(y - y_0) \tag{3.14}$$

donde A y (x_0, y_0) representan el área y las coordenadas del centroide de la lesión, respectivamente. A partir de estos momentos de segundo orden es posible determinar los coeficientes (a,b y c) de la elipse como:

$$\begin{bmatrix} a & b \\ b & c \end{bmatrix} = \frac{1}{4 \cdot s_{xx} \cdot s_{yy} - s_{xy}^2} \cdot \begin{bmatrix} s_{yy} & -s_{xy} \\ -s_{xy} & s_{xx} \end{bmatrix} \tag{3.15}$$

La elipse equivalente puede ser usada para describir de manera burda la forma de la lesión. Para cualquier píxel $P(x_1, y_1)$ en el borde de la lesión existe un píxel de cruce $C(x_2, y_2)$ en la elipse equivalente el cual es encontrado mediante una línea que va desde el centro de la elipse equivalente hasta $P(x_1, y_1)$.

La disimilaridad entre $P(x_1, y_1)$ y $C(x_2, y_2)$ puede ser calculada como [39]:

$$D(P) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \tag{3.16}$$

Finalmente, es posible determinar el descriptor de **Forma** como:

$$F = \frac{\sum_{P \in LB} D(P)}{BL_N} \tag{3.17}$$

donde BL_N es el número de píxeles en el borde de la lesión.

Es posible medir el ángulo de una lesión mediante el ángulo del eje mayor (θ) de su elipse equivalente. Sin embargo, para determinar la orientación de una elipse es necesario cuantificar su grado de paralelismo (en relación al transductor US) en el rango $[0, \frac{\pi}{2}]$. Por tanto, el parámetro de orientación se obtiene ajustando el ángulo(θ) en el rango $[0, \frac{\pi}{2}]$ como [39]:

$$O_E = \begin{cases} \theta & \text{if } 0 \leq \theta \leq \frac{\pi}{2} \\ \pi - \theta & \text{if } \frac{\pi}{2} \leq \theta \leq \pi \\ \theta - \pi & \text{if } \pi \leq \theta \leq \frac{3\pi}{2} \\ 2\pi - \theta & \text{if } \frac{3\pi}{2} \leq \theta \leq 2\pi \end{cases} \tag{3.18}$$

La **anfractuosidad** de una lesión se define como la proporción que existe entre la lesión y su elipse equivalente [38], y se denota como:

$$ENC = \frac{N_L}{N_E} \tag{3.19}$$

donde N_L y N_E denotan el número de píxeles en el borde de la lesión y la elipse equivalente, respectivamente.

Otro parámetro que describe la forma de la lesión es la proporción **eje mayor-eje menor** de la elipse equivalente [38], que se define como:

$$L : S = \frac{L}{S} \tag{3.20}$$

donde L y S son las longitudes del eje mayor y eje menor de la elipse equivalente, respectivamente.

3.2.1.4. Mapa de distancias

Para cualquier píxel $P(x, y)$ en una imagen que contiene una lesión previamente segmentada, la distancia de $P(x, y)$ al borde de la lesión se define de manera recursiva como [39]:

$$Distancia(P) = \text{Min}\{Distancia(N_8(P))\} + 1 \tag{3.21}$$

donde la distancia para cualquier píxel en el borde de la lesión es igual a cero y $N_8(P)$ es el vecindario de P definido como:

$$\begin{aligned} N_8(P) = \{ &\text{ (x-1, y-1), (x, y-1), (x+1, y-1), (x-1, y),} \\ &\text{(x+1, y), (x-1, y+1), (x, y+1), (x+1, y+1) }\} \end{aligned} \tag{3.22}$$

Un máximo círculo inscrito es encontrado a partir del mapa de distancias, el cual divide algunas

áreas lobuladas como se muestra en la Figura 4.6. Es razonable cuantificar la irregularidad del contorno mediante el **número de lobulaciones notables** [39], para ello es necesario establecer un umbral U de manera que las áreas menores que dicho umbral serán descartadas.

El mapa de distancias puede ser usado para determinar una región de tejido circundante a la lesión, así como una región interna de la lesión adyacente a su borde, como se observa en la Figura 4.7. La intensidad media de niveles de gris para ambas regiones con una anchura k a partir del borde de la lesión se definen como:

$$avg_{Circ} = \frac{\sum_{Distancia(P)=1}^{k} I(P)}{N_{Circ}} \quad \text{y} \quad avg_{Int} = \frac{\sum_{Distancia(P)=1}^{k} I(P)}{N_{Int}} \tag{3.23}$$

donde $I(P)$ es la intensidad de gris para el píxel P, y N_{Circ} y N_{Int} representa el número de píxeles para la región de tejido circundante y la región interna, respectivamente. Entonces se puede evaluar el **grado de interfaces abruptas** como [39]:

$$LB_D = avg_{Circ} - avg_{Int} \tag{3.24}$$

La intensidad media de niveles de gris es también utilizada para cuantificar el patrón del eco como [39]:

$$EP_I = \frac{\sum_{P \in Int} I(P)}{N_{Int}} \tag{3.25}$$

donde $I(P)$ es la intensidad de gris para un píxel P dentro de la región interna del tumor y N_{Int} denota el número de píxeles dentro de la región del tumor.

Por último, es posible cuantificar el **patrón interno del eco** mediante la magnitud del gradiente

como [39]:

$$EP_{AG} = \frac{\sum_{P \in Mass} G(P)}{N_{Mass}}, \tag{3.26}$$

donde $G(P)$ representa la magnitud del gradiente en P que se puede computar por medio de filtros de Sobel [40].

3.2.1.5. Esqueleto

El esqueleto es un representación efectiva de una región, frecuentemente utilizada en áreas como reconocimiento de patrones y visión por computadora. Sea B_R el conjunto de puntos en el borde de una región R, el esqueleto de R es el conjunto de puntos $x \in X$, donde x está dentro de R y existen al menos dos puntos p_i y p_j, en B_R tales que:

$$d(x, p_i) = d(x, p_j) = \min\{d(x, p_k) | p_k \in B_R\}$$

donde $d(\cdot)$ es una métrica de distancia.

Mientras mayor es el número de protuberancias y depresiones de una lesión más complejo se vuelve el esqueleto [38]. Dos parámetros son calculados a fin de cuantificar la complejidad del esqueleto:

1. Si bien parece lógico medir la complejidad del esqueleto mediante el número de puntos que contiene, cabe recordar que esta característica es dependiente del tamaño de la lesión, para ello se propone el **esqueleto elíptico-normalizado**, que se define mediante [38]:

$$ENS = \frac{N_{SK}}{P_E} \tag{3.27}$$

donde N_{SK} es el número de puntos del esqueleto y P_E es el perímetro de la elipse equivalente del tumor.

2. Otro parámetro que describe la complejidad del esqueleto es el **número de puntos terminales**, como se muestra en la Figura 4.9.

3.2.1.6. Geométricos

La geometría es un factor importante en la correcta clasificación de lesiones de mama de ultrasonografía. A continuación se describen cuatro descriptores basados en esta característica:

1. **Circularidad [37]:**

$$C = \frac{P^2}{Area(S)}, \tag{3.28}$$

donde P es el perímetro de la lesión S.

2. **Cociente morfológico [22]:**

$$msahpe = \frac{Area(S)}{Area(S_c)}, \tag{3.29}$$

donde S representa la lesión y S_c la cerradura morfológica de S con un elemento estructurante circular de 10 píxeles de radio.

3. **Relación profundidad-anchura [38]:**

$$D : W = \frac{D}{W} \tag{3.30}$$

donde D y W denotan las longitudes de los bordes horizontal y vertical del mínimo rectángulo que contiene a la lesión.

4. **Tamaño.** Es posible evaluar ésta característica mediante el número de píxeles que contiene la lesión.

3.2.2 Descriptores de textura

Se ha demostrado que la información de textura es un factor importante en la clasificación de lesiones de mama de ultrasonografía. En un sistema CAD una textura se interpreta como la variación en los niveles de gris a escalas menores que las escalas de interés. A continuación se describen las diferentes técnicas utilizadas para cuantificar la heterogeneidad en los niveles de gris.

3.2.2.1. Curva de complejidad

Para generar la curva de complejidad (CC) se aplican distintos valores de umbral (α) a la imagen bidimensional $f(x, y)$ de USM, para crear diferentes imágenes binarias como [14]:

$$f^{\alpha}(x,y) = \begin{cases} 1, & \forall f(x,y) \geq \alpha \\ 0, & \forall f(x,y) < \alpha \end{cases}$$

donde $\alpha = \{0, 1, ..., G-1\}$, G es el valor máximo de niveles de gris en la imagen. La CC está relacionada con el número de transiciones de 0 a 1 como:

$$C(\alpha) = \frac{T^{\alpha}_{1,0} + T^{\alpha}_{0,1}}{N_x \times (N_y - 1) + N_y \times (N_x - 1)}$$

donde $T^{\alpha}_{1,0}$ y $T^{\alpha}_{0,1}$ representan el número total de transiciones en $f^{\alpha}(x, y)$ en las direcciones vertical y horizontal, respectivamente, y $N_x \times N_y$ es el número de filas y columnas en $f(x, y)$.

Cinco descriptores son extraídos a partir de la curva de complejidad [14], como se muestra en la Tabla 3.1.

Nombre	Ecuación
Valor medio de transiciones	$vmed = \frac{\sum_{\alpha=1}^{G-1} C(\alpha)}{G-1}$
Valor máximo de transiciones	$vmax = \max[C(\alpha)] \quad 0 \leq \alpha \leq G-1$
Media de la muestra	$mmed = \frac{\sum_{\alpha=1}^{G-1} \alpha \cdot C(\alpha)}{\sum_{\alpha=1}^{G-1} C(\alpha)}$
Desviación estándar de la muestra	$mstd = \sqrt{\frac{1}{\sum_{\alpha=0}^{G-1} C(\alpha)} \cdot \sum_{\alpha=0}^{G-1} (\alpha - mmed) \cdot C(\alpha)}$
Entropía	$ent = \sum_{\alpha=0}^{G-1} C(\alpha) \cdot \log[C(\alpha)], \quad C(\alpha) > 0$

Tabla 3.1: Descriptores extraídos de la curva de complejidad.

3.2.2.2. Matriz de co-ocurrencia de los niveles de gris

Para una imagen con G niveles de gris, la matriz de co-ocurrencia (GLCM, por sus siglas en inglés) es un histograma bidimensional $G \times G$ a partir de las probabilidades de ocurrencia $p(i, j)$, lo cual se define como el número de ocurrencias de píxeles-pares (i, j) donde un nivel i está espaciado hacia un nivel j por una distancia d a lo largo de una dirección θ [14]. Esto se muestra en la Figura 3.1 para una imagen $f(x, y)$ con dimensiones 3×5, ocho niveles de gris, $d = 1$ y $\theta = 0°$.

Veintidós descriptores de textura pueden ser calculados a partir de la GLCM, como se muestra en la Tabla 3.2.

Descriptor	Ecuación	Referencia
Auto-correlación	$\sum_i \sum_j (i \cdot j) p(i, j)$	[41]
Contraste	$\sum_i \sum_j \lvert i - j \rvert^2 p(i, j)$	[41]
Correlación I	$\sum_i \sum_j \frac{(i - \mu_x)(j - \mu_y) p(i, j)}{\sigma_x \sigma_y}$	[42]

Correlación II	$\sum_i \sum_j \frac{(i \cdot j)p(i,j) - (\mu_x \mu_y)}{\sigma_x \sigma_y}$	[41]
Agrupamiento de protuberancias	$\sum_i \sum_j (i + j - \mu_x - \mu_y)^4 p(i,j)$	[41]
Agrupamiento de sombras	$\sum_i \sum_j (i + j - \mu_x - \mu_y)^3 p(i,j)$	[41]
Disimilaridad	$\sum_i \sum_j \lvert i - j \rvert \cdot p(i,j)$	[41]
Energía	$\sum_i \sum_j p(i,j)^2$	[41]
Entropía	$-\sum_i \sum_j p(i,j) \cdot \ln(p(i,j))$	[41]
Homogeneidad I	$\sum_i \sum_j \frac{p(i,j)}{1 + \lvert i - j \rvert}$	[41]
Homogeneidad II	$\sum_i \sum_j \frac{p(i,j)}{1 + \lvert i - j \rvert^2}$	[41]
Máxima probabilidad	$\max_{i,j} p(i,j)$	[41]
Suma de cuadrados	$\sum_i \sum_j (i - v)^2 p(i,j)$	[43]
Suma de promedios	$\sum_{i=2}^{2L} i \cdot p_{x+y}(i)$	[43]
Suma de entropía **(SE)**	$-\sum_{i=2}^{2L} p_{x+y}(i) \cdot \ln(p_{x+y}(i))$	[43]
Suma de varianza	$\sum_{i=2}^{2L} (i - \mathbf{SE}) \cdot p_{x+y}(i)$	[43]
Diferencia de varianza	$\sum_{i=0}^{L-1} i^2 \cdot p_{x-y}(i)$	[43]
Diferencia de entropía	$-\sum_{i=0}^{L-1} p_{x-y}(i) \cdot \ln(p_{x-y}(i))$	[43]
Medida de información de correlación I	$\frac{H(X,Y) - H_1(X,Y)}{\max(H(X), H(Y))}$	[43]
Medida de información de correlación II	$\sqrt{1 - \exp\left[-2(H_2(X,Y) - H(X,Y))\right]}$	[43]
Diferencia inversa normalizada	$\sum_i \sum_j \frac{p(i,j)}{1 + \lvert i - j \rvert^2 / L}$	[44]
Desviación estándar	$\sqrt{\frac{1}{G^2} \sum_{i,j} (p(i,j) - \overline{p})^2}$	[14]

Tabla 3.2: Descriptores de textura extraídos de la GLCM.

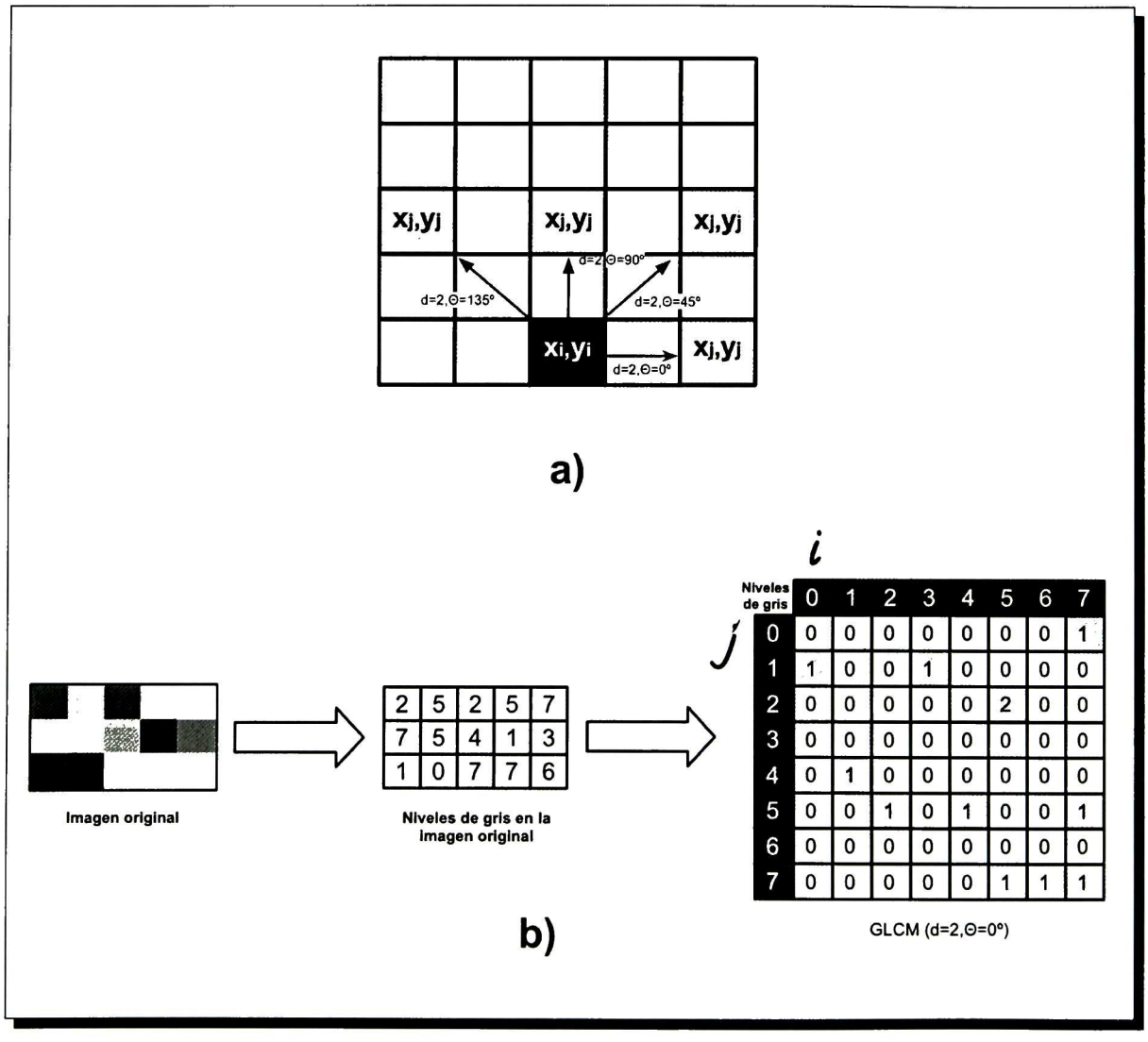


Figura 3.1: (a) Relación espacial entre dos píxeles para las cuatro distintas orientaciones $\theta = 0°, 45°, 90°, 135°$ con una distancia $d = 2$. (b) Ejemplo de la GLCM con orientación $\theta = 0°$, distancia $d = 1$ para una imagen con $G = 8$ niveles de gris.

En la Tabla 3.3 se describen los términos utilizados en la extracción de características de textura basados en la GLCM.

3.2.2.3. *Coeficientes de auto-correlación y auto-covarianza*

Los coeficientes de autocorrelación entre $P_1(i,j)$ y $P_2(i+\Delta m, j+\Delta n)$ dentro de una imagen de tamaño $M \times N$ se definen como [13]:

$$A(\Delta m, \Delta n) = \frac{1}{(M-\Delta m)(N-\Delta n)} \times \sum_{x=0}^{M-1-\Delta m} \sum_{y=0}^{N-1-\Delta n} f(x,y) f(x+\Delta m, y+\Delta n) \tag{3.31}$$

Los coeficientes de autocorrelación tienen la desventaja de ser sensibles al brillo de la imagen, para ello se proponen los coeficientes de autocovarianza con media cero, que se definen como [13]:

$$A(\Delta m, \Delta n) = \frac{1}{(M-\Delta m)(N-\Delta n)} \times \sum_{x=0}^{M-1-\Delta m} \sum_{y=0}^{N-1-\Delta n} (f(x,y)-\overline{f})(f(x+\Delta m, y+\Delta n)-\overline{f}) \tag{3.32}$$

donde $\overline{f}$ es el valor medio de $f(x,y)$.

3.2.2.4. *Propiedades por bloques*

- **Bloque de diferencia inversa de probabilidades** $(BDIP)$. Se define como la diferencia entre el número de píxeles de un bloque de la siguiente manera [45]:

$$BDIP = P^2 - \frac{\sum_{x,y \in B} f(x,y)}{\max_{(x,y)\in B} f(x,y)}, \tag{3.33}$$

Término	Definición
$p(i,j)$	Probabilidad del elemento (i,j) en la GLCM
N_g	Número de niveles de gris
$p_x(i)$	Probabilidad marginal de la suma de probabilidades de los renglones para la columna i, $= \sum_{j=1}^{N_g} p(i,j)$
$p_y(j)$	Probabilidad marginal de la suma de probabilidades de las columnas para el renglón j, $= \sum_{i=1}^{N_g} p(i,j)$
μ_x, μ_y	Son las medias de p_x y p_y, respectivamente
σ_x, σ_y	Son las desviaciones estándar de p_x y p_y, respectivamente
$p_{x+y}(k)$	$\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p(i,j), k = 2,3,\ldots,2N_g$
$p_{x-y}(k)$	$\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p(i,j), k = 0,1,\ldots,N_g-1$
$H(X)$ y $H(Y)$	Son la entropías de p_x y p_y respectivamente
$H(X,Y)$	$-\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} P(i,j)\ln(P(i,j))$
$H_1(X,Y)$	$-\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} P(i,j)\ln(p_x(i)p_y(j))$
$H_2(X,Y)$	$-\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p_x(i)p_y(j)\ln(p_x(i)p_y(j))$

Tabla 3.3: Descripción de términos utilizados para el cálculo de los descriptores de textura extraídos de la GLCM.

donde B denota un bloque de tamaño $P \times P$. Mientras más grande es la variación de intensidades en un bloque, mayor es el valor $BDIP$.

- **Bloque de variación local de coeficientes de correlación** $(BVLC)$. Mide la suavidad de la textura. El valor $BVLC$ se define como [45]:

$$BVLC = \max_{(k,l)\in O_4} [\rho(k,l)] - \min_{(k,l)\in O_4} [\rho(k,l)], \tag{3.34}$$

donde $O_4 = \{(0,1),(1,0),(1,1),(1,-1)\}$, y

$$\rho(k,l) = \frac{(1/P^2)\sum_{(x,y)\in B} f(x,y)f(x+k,y+l)\mu_{0,0}\mu_{k,l}}{\sigma_{0,0},\sigma_{k,l}} \tag{3.35}$$

donde $\mu_{0,0}$ y $\sigma_{0,0}$ representan el valor medio y la desviación estándar del bloque de tamaño $P \times P$. El término (k,l) denota cuatro orientaciones $(-90°, 0°, -45°, 45°)$. Como resultado, $\mu_{k,l}$ y $\sigma_{k,l}$ representan el valor medio y la desviación estándar del bloque recorrido, respectivamente.

3.3 Selección de características

En el área de reconocimiento de patrones, la selección de características es un término que se usa para describir la tarea de reducir las entradas de un clasificador a un tamaño apropiado para su procesamiento y análisis, de manera que el desempeño de clasificación pueda aumentar.

3.3.1 Análisis de componentes principales

El análisis de componentes principales es una técnica utilizada para reducir la dimensionalidad de un conjunto de datos. A fin de eliminar las relaciones entre variables correlacionadas (que miden información común) es posible transformar el espacio original de características en otro conjunto de nuevas variables incorreladas entre sí (sin repetición o redundancia en la información) llamado espacio

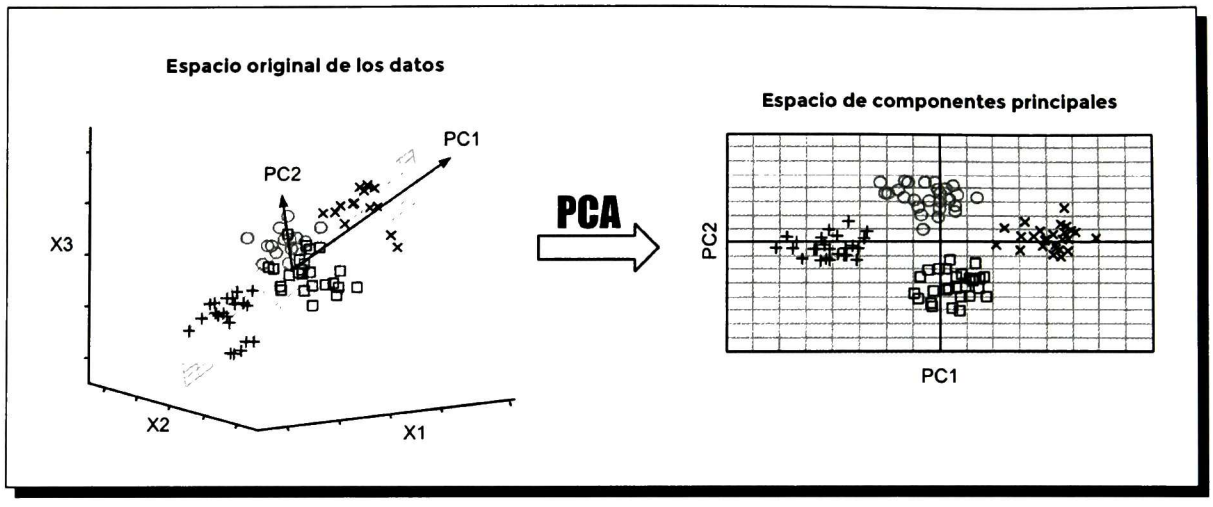


Figura 3.2: Ejemplo de análisis de componentes principales, donde X1, X2 y X3 representan las características en el espacio original, y PC1 y PC2 representan los componentes principales

de componentes principales, donde las nuevas variables son combinaciones lineales de las originales y se construyen según el orden de importancia en cuanto a la variabilidad total que recogen de las muestras. En la Figura 3.2 se observa un ejemplo de proyección de los datos mediante PCA.

El PCA parte de un conjunto n de muestras, donde cada una tiene M atributos, y cuyo objetivo es que cada una de esas muestras se describa con sólo m componentes principales, donde $m < M$, es decir, el número de componentes principales m tiene que ser inferior al número total de atributos en el espacio de características X.

El PCA se basa en la descomposición en vectores propios de la matriz de covarianza, la cual se calcula con la siguiente ecuación [46]:

$$\text{cov}(X) = \frac{X^T X}{n-1}$$

$$\text{cov}(X)\ p_a = \lambda_a\ p_a$$

$$\sum_{a=1}^{M} \lambda_a = 1$$

donde λ_a es el valor propio asociado al vector propio p_a y los valores propios λ_a miden la cantidad de varianza capturada por cada PC. Finalmente se realiza la proyección del espacio original X en p_a como: $t_a = Xp_a$, donde t_a son las proyecciones de X en p_a.

Es posible entonces aproximar el conjunto original de datos como:

$$X = \sum_{a=1}^{m} t_a p_a^T + E$$

donde t_a son vectores conocidos como *scores* y contienen la información de cómo las muestras están relacionadas unas con otras, los vectores p_a se llaman *loadings* e informan de la relación existente entre las variables y E es una matriz de error que se produce al utilizar pocos PCs.

3.3.2 Criterio de mínima-redundancia-máxima-relevancia (mRMR)

El criterio mRMR tiene como objetivo encontrar a partir de un espacio M-dimensional de características, R^M, un subespacio de m característcias, R^m, $(m < M)$ que genere la mayor relevancia con la clase c deseada y menor redundancia entre ellas [47].

Supóngase un subespacio de m características $S = \{x_i, i = 1, \cdots, m\}$. El criterio mRMR está dividido en dos fases:

1. La primera fase consiste en encontrar las características que maximicen la relevancia (D) con la clase c de manera que se satisfaga la siguiente ecuación:

$$\text{max} D(S, c), D = \frac{1}{m} \sum_{x_i \in S} (H(x_i) + H(c) - H(x_i, c)) \tag{3.36}$$

donde $H(x_i)$ y $H(c)$ son las entropías de la característica i y de la clase c, respectivamente y $H(x_i, c)$ representa la entropía conjunta de i y c.

2. La segunda fase consiste en minimizar la redundancia (R) entre las características seleccionadas con máxima relevancia, de tal manera que si dos variables presentan alta dependencia, el poder discriminativo no disminuirá si se excluye alguna de ellas y se define como:

$$\min R(S), R = \frac{1}{m^2} \sum_{x_i \in S} (H(x_i) + H(x_j) - H(x_i, x_j)) \tag{3.37}$$

donde $H(x_i)$ y $H(x_j)$ son las entropías de las características i y j, respectivamente y $H(x_i, x_j)$ representa la entropía conjunta de las características i y j.

Finalmente se combinan ambas condiciones. Para ello se emplea el operador $\max \Phi(D, R)$, cuyo máximo optimiza de manera simultánea D y R como:

$$\Phi(D, R), \Phi = D - R \tag{3.38}$$

Es posible utilizar un método incremental para determinar el conjunto de características $\Phi(D, R)$. Supóngase un conjunto de datos de entrada con M características $X = \{x_1, x_2, \ldots, x_M\}$, una variable de clase c, y un subconjunto de m características $S = \{x_1, x_2, \ldots, x_m\}$. El objetivo es determinar la m-ésima característica del conjunto $\{X - S_{m-1}\}$, para ello se deberá maximizar $\Phi(\cdot)$. El algoritmo incremental optimiza la Ecuación 3.38 como:

$$\max_{x_j \in X - S_{m-1}} \left[I(x_j; c) - \frac{1}{m-1} \sum_{x_j \in S_{m-1}} I(x_j; x_i) \right] \tag{3.39}$$

3.4 Análisis lineal discriminante de Fisher

El análisis lineal discriminante de Fisher (FLDA, por sus siglas en inglés) es una técnica estadística que puede emplearse para realizar la separación de un conjunto de datos en dos o más clases; mediante la proyección de los datos de entrada en una dirección determinada, como se observa en la Figura 3.3 . Dicho de otra manera, el FLDA es una transformación de un espacio de entrada d-dimensional a un espacio de salida uno-dimensional, donde el número de clases $C = 2$ [48]. Para cada elemento $X = \{x_1, x_2, \ldots, x_n\}$ se realiza la proyección $Y = \{y_1, y_2, \ldots, y_n\}$ en la dirección w como:

$$y_i = w^T x_i,\ i = 1, 2, \ldots, n \tag{3.40}$$

Los valores de dirección w se calculan como:

$$w = S_W^{-1}(\mu_1 - \mu_2) \tag{3.41}$$

donde μ_1 y μ_2 son las medias de los datos para la clase 1 y la clase 2, respectivamente y S_W es la matriz de covarianza intra-clase, que se define como:

$$S_W = \sum_{i=1}^{C} E[(X - \mu_i)(X - \mu_i)^T] \tag{3.42}$$

Finalmente, la clasificación de un patrón desconocido, x, se define como:

$$D(x) = \left[x - \frac{1}{2}(\mu_1 + \mu_2)\right]^T w \tag{3.43}$$

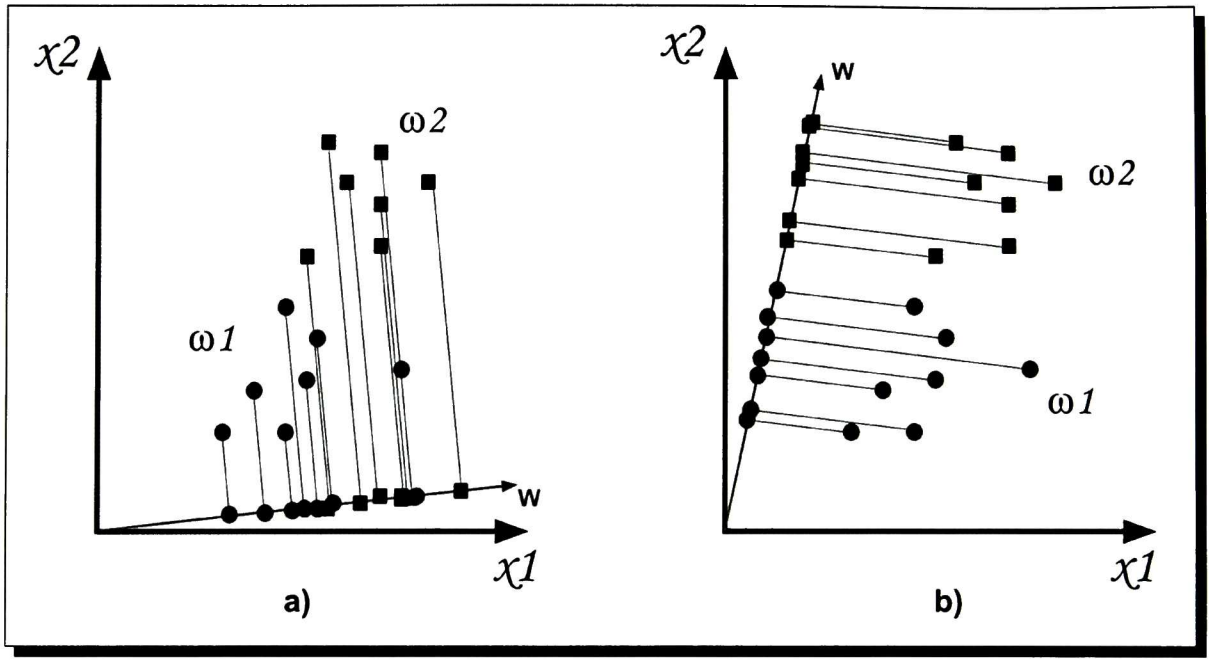


Figura 3.3: Ejemplo de (a) una mala proyección y (b) una buena proyección

3.5 Estimación del error

El conjunto de entrenamiento $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ consiste de $\mathbf{n}$ observaciones $x_i = (t_i, y_i)$, donde t_i es un vector de características p-dimensional y y_i es su clase o respuesta verdadera. Entonces, dada un regla de predicción $r_{\mathbf{x}}(t)$ se desea estimar la tasa de error de dicha regla en la predicción de respuestas futuras. La notación $Q[y, r]$ indica la discrepancia entre el valor predicho por la regla r y la respuesta verdadera y. En el caso de una situación dicotómica (dos clases) la discrepancia Q se expresa como [49]:

$$Q[y, r] = \begin{cases} 0, & \text{si } r = y \\ 1, & \text{si } r \neq y \end{cases} \tag{3.44}$$

3.5.1 Método por resustitución

La tasa de error aparente (o método por resustitución, RSB) utiliza un mismo conjunto $\mathbf{x}$ tanto para el entrenamiento como para la evaluación del desempeño, que se estima como:

$$\hat{e}_n^{RSB} = \frac{1}{n}\sum_{i=1}^{n} Q[y_i, r_x(t)] \tag{3.45}$$

El método por resustitución no es adecuado para la evaluación del desempeño de un clasificador, ya que tiende a subestimar el error debido al traslape entre los conjuntos de entrenamiento y prueba.

3.5.2 Método *bootstrap*

Para superar la limitación de pocas muestras disponibles, los métodos *bootstrap* crean artificialmente nuevos datos para poder evaluar con mayor precisión las propiedades estadísticas de un estimador de error. Sea $\mathbf{x}$ el conjunto de datos disponibles de tamaño $\mathbf{n}$. Entonces, el conjunto de entrenamiento *bootstrap*, $\mathbf{x}^*$, de tamaño $\mathbf{n}$, se forma aleatoriamente tomando muestras con remplazo del conjunto $\mathbf{x}$. Es decir que una muestra en particular, x_i, puede copiarse varias veces al conjunto $\mathbf{x}^*$.

El método *bootstrap* ordinario consiste en crear un conjunto $\mathbf{x}^*$ para entrenar una regla de predicción r, evaluando el desempeño con el conjunto original $\mathbf{x}$. Esto no es muy conveniente debido a que existe traslape entre el conjunto de entrenamiento y prueba. Para superar este inconveniente, el método *bootstrap* dejando uno fuera (LOOBS), genera un total de B conjuntos *bootstrap* de tamaño $\mathbf{n}$. Cada muestra observada se clasifica repetidamente usando los conjuntos *bootstrap* en los cuales esa muestra en particular no aparece. De esta manera el método evita clasificar muestras que fueron utilizadas para construir el modelo de predicción.

Sea N_i^b el número de veces que la muestra x_i se incluye en el b-ésimo conjunto *bootstrap*, tal que:

$$I_i^b = \begin{cases} 1 & \text{si } N_i^b = 0 \\ 0 & \text{si } N_i^b > 0 \end{cases} \tag{3.46}$$

de modo que cuando $I_i^b = 1$ significa que la muestra x_i no está contenida en $\mathbf{x}^{*b}$. También se define la discrepancia como:

$$Q_i^b = Q[y_i, r_{\mathbf{x}^{*b}}(t_i)] \tag{3.47}$$

Entonces el error de LOOBS se estima como:

$$\hat{e}_n^{LOOBS} = \frac{1}{n}\sum_{i=1}^{n} \hat{E}_i \tag{3.48}$$

donde

$$\hat{E}_i = \frac{\sum_b I_i^b Q_i^b}{\sum_b I_i^b} \tag{3.49}$$

3.5.3 Estimador *bootstrap*.632+

Se estima que un conjunto *bootstrap* tiene aproximadamente .632$\mathbf{n}$ muestras únicas del conjunto original de datos [50]. Esto se deriva de la probabilidad de que $x_0 = x_i$ aparezca en el conjunto entrenamiento es $1 - (1 - 1/n^2) \approx .632$. Por tanto, el inconveniente del método LOOBS es que produce una predicción del error sobreestimada. Para corregir esto se ha sugerido el estimador *bootstrap* .632+, que se define como [49]:

$$\hat{e}_n^{.632+} = w\hat{e}_n^{LOOBS} + (1 - w)\hat{e}_n^{RSB} \tag{3.50}$$

donde el peso de penalización w está en función de la tasa de traslape relativo R y la tasa de error de no información γ, definidos como:

$$w = \frac{.632}{1 - .368R} \tag{3.51}$$

$$R = \frac{\hat{e}_n^{LOOBS} - \hat{e}_n^{RSB}}{\gamma - \hat{e}_n^{RSB}} \tag{3.52}$$

$$\gamma = \sum_{i=1}^{n} \sum_{j=1}^{n} Q[y_i, r_x(t_j)]/n^2 \tag{3.53}$$

3.6 Conclusiones

En este capítulo se mostró el fundamento teórico de las técnicas implementadas para el desarrollo de esta tesis. Primeramente se describen las diferentes técnicas de descripción de lesiones de mama de US. El objetivo de estas técnicas es cuantificar de manera numérica la irregularidad del contorno de la lesión, así como la variación en sus niveles de gris. Posteriormente se definieron las técnicas de PCA y MI, las cuales buscan reducir la dimensionalidad del espacio de características, preservando la mayor cantidad de información discriminante. Así mismo se describió la técnica de clasificación FLDA, la cual se basa en la proyección de los datos de entrada en una dirección determinada. Por último se describió la técnica de estimación del error *Bootstrap*.632+.

4

Metodología

4.1 Introducción

En este capítulo se describe la metodología seguida en el desarrollo de esta tesis. Para ello se han preparado seis secciones: en la primera se describe el algoritmo de segmentación utilizado, en las secciones dos y tres se describe el proceso de extracción de características morfológicas y de textura, en la cuarta sección se presenta el método de normalización de los datos para su análisis, en la sección cinco se definen los espacios de características construidos y, finalmente, en la sección seis se describe el proceso de selección de atributos.

En la Figura 4.1 se presenta el diagrama de bloques general de la metodología seguida en esta investigación.

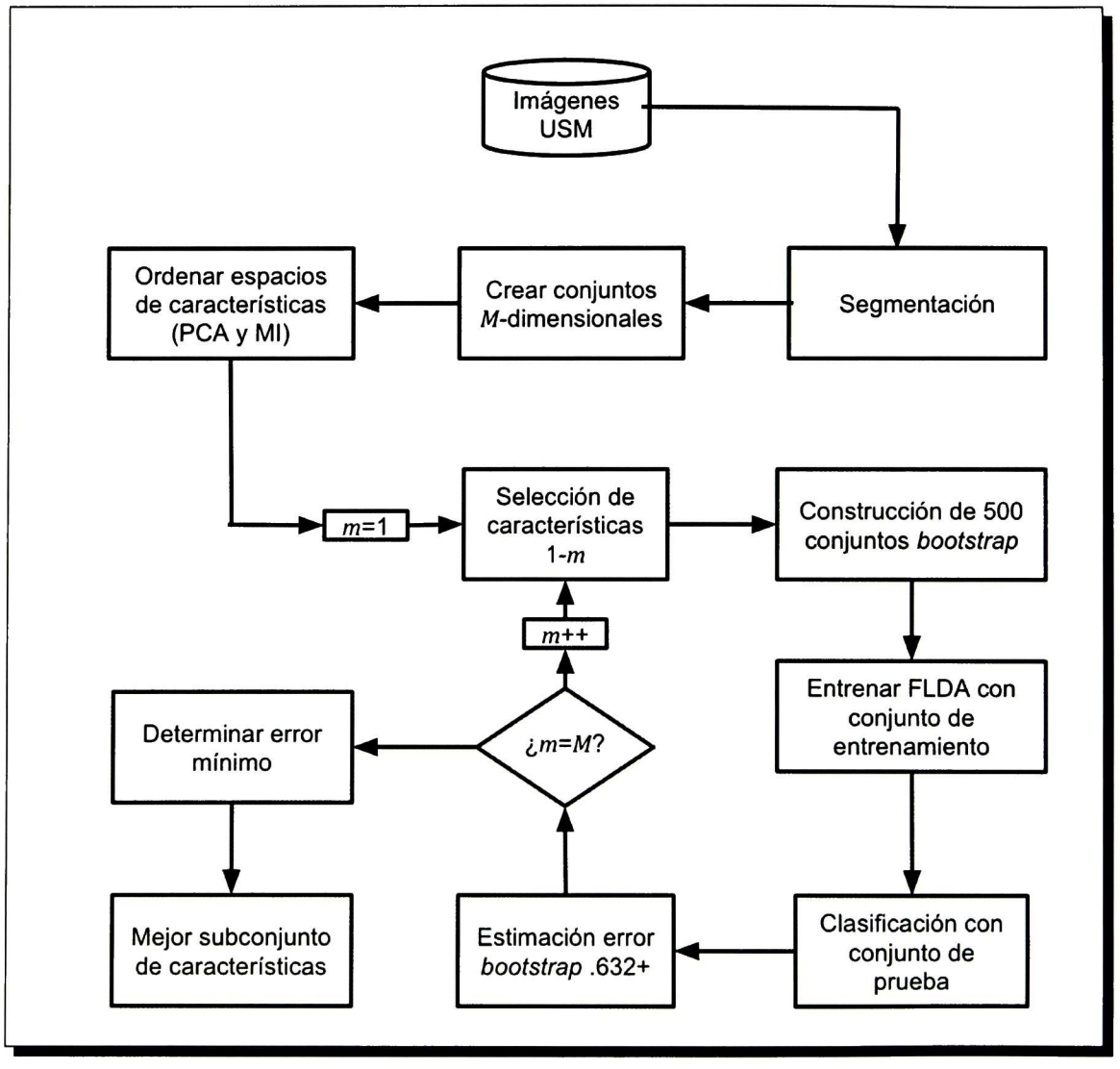


Figura 4.1: Diagrama a bloques de la metodología.

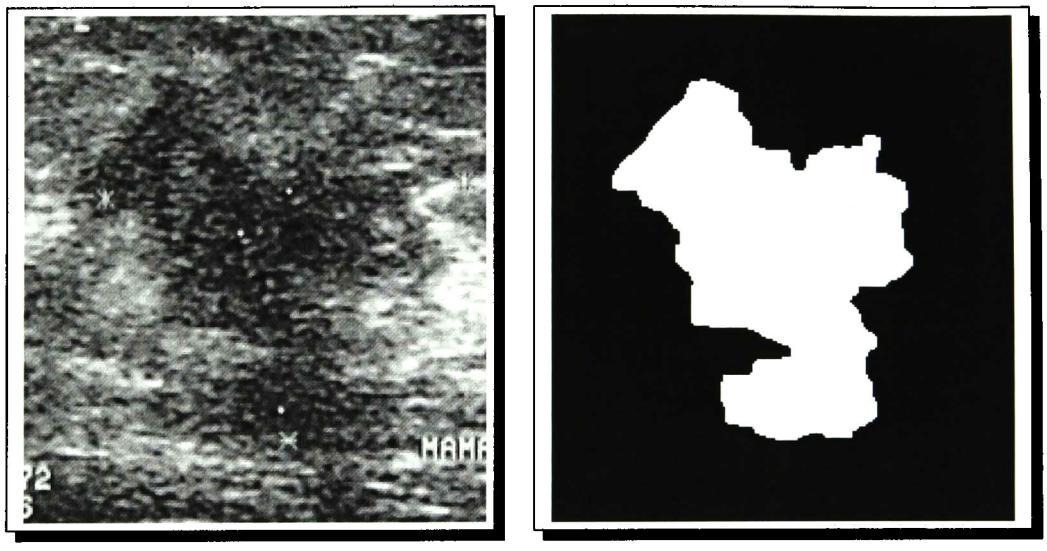
4.2 Segmentación

Se contó con un banco de 638 imágenes de USM (413 lesiones benignas y 225 carcinomas) adquiridas durante rutinas de diagnóstico en el Instituto Nacional del Cáncer (INCA) de Río de Janeiro, Brazil. El comité de ética e investigación del INCA aprobó el uso de éstas imágenes mediante el convenio nº 38/2001. Dichas imágenes fueron obtenidas mediante un equipo Sonoline Sienna a 7.5 MHz utilizando un transductor con arreglo lineal B-mode de 40mm. Las imágenes fueron capturadas directamente de la señal de video de 8-bits (con 256 niveles de gris) y guardadas en formato TIFF.

Todas las imágenes fueron segmentadas mediante un método basado en la transformada *watershed* controlada por marcadores [11]. Primeramente se normaliza el rango de intensidad de la imagen de 0 a 255 y se aplica una técnica de mejoramiento del contraste mediante la técnica CLAHE[1] para ecualización del histograma. Posteriormente se emplea un filtro de difusión anisotrópico guiado por descriptores de textura a fin de reducir el artefacto *speckle*. Para eliminar los píxeles distantes que no pertenecen a la lesión se aplica una función de restricción Gaussiana. Después, se realiza un proceso de umbralado iterativo para generar distintas imágenes binarias, las cuales son utilizadas para construir funciones marcadoras y así utilizar la transformada *watershed* para crear potenciales bordes de la lesión.

Como resultado del algoritmo de segmentación se obtiene una imagen binaria generada a partir de una imagen de USM en escala de grises, como se observa en la Figura 4.2.

[1] *Contrast Limited Adaptive Histogram Equalizations.*



(a) Imagen de USM

(b) Imagen binaria donde la región en negro representa el fondo y la región blanca representa la lesión

Figura 4.2: Resultado del método de segmentación utilizado.

4.3 Extracción de características morfológicas

La extracción de características morfológicas se realizó a partir de la imagen binaria obtenida del proceso de segmentación (ver Figura 4.2). Para ello se implementaron diferentes técnicas de descripción propuestas en la literatura especializada.

4.3.1 Longitud radial normalizada

La NRL mide la distancia que existe entre el centroide de la lesión y cada uno de los píxeles del contorno. Para determinar el centroide de la lesión se utilizaron los momentos geométricos de la imagen binaria como:

$$\overline{x} = \frac{m_{10}}{m_{00}} \quad y \quad \overline{y} = \frac{m_{00}}{m_{00}}, \tag{4.1}$$

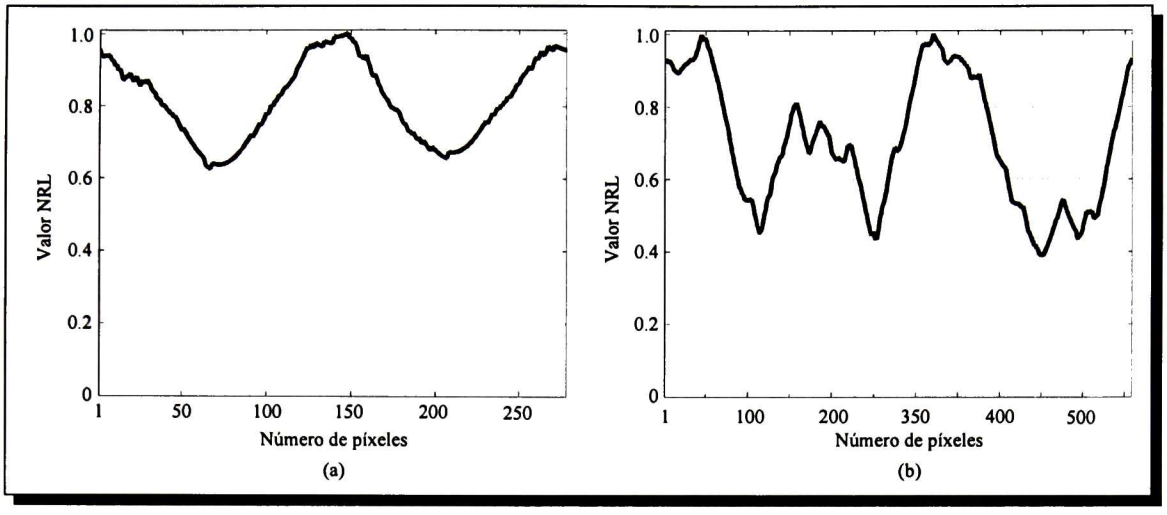


Figura 4.3: Señal obtenida de la NRL para (a) una lesión y benigna y (b) un carcinoma.

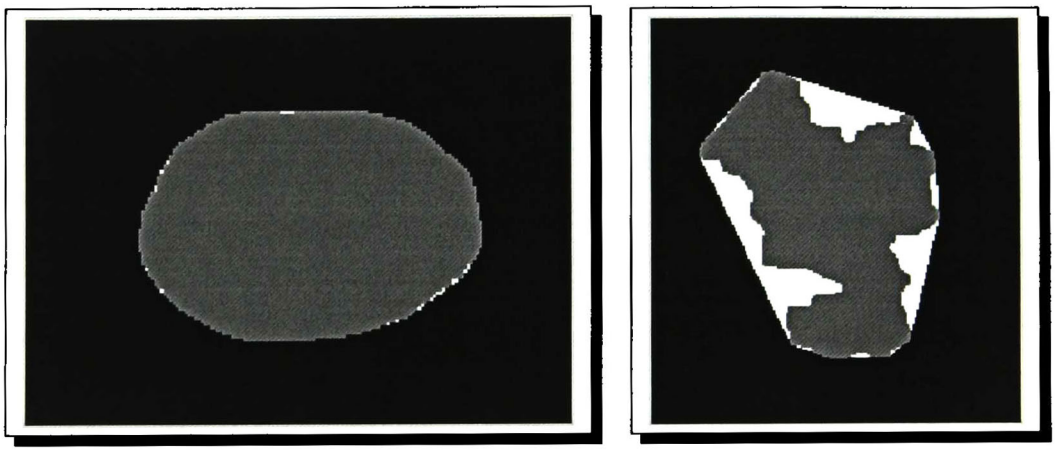
donde $(\overline{x}, \overline{y})$ son las coordenadas del centroide de la lesión y m_{pq} son los momentos de orden $p+q$ definidos como:

$$m_{pq} = \sum_x \sum_y f(x,y) \tag{4.2}$$

Posteriormente se calculó la distancia Euclidiana entre el centroide de la lesión y cada píxel de su contorno, obteniendo como resultado una señal 1-dimensional (ver Figura 4.3) de la cual se calcularon los parámetros: desviación estándar, razón de área, entropía y rugosidad del contorno, definidos en el Capítulo 3, a fin de caracterizar dicha señal.

4.3.2 Envolvente convexa

Primeramente se determina la envolvente convexa (EC) de la lesión, como se observa en la Figura 4.4 y posteriormente se calcularon algunas operaciones de conjuntos para computar los atributos de relación de áreas y valor residual normalizado con el objetivo caracterizar la irregularidad de la lesión, como se define en el Capítulo 3.



(a) Lesión benigna (b) Carcinoma

Figura 4.4: Representación de la EC (área en blanco) para una lesión (área en gris)

Se determinaron puntos cóncavos y convexos como se define en el Capítulo 3 con el objetivo de dividir la lesión en regiones lobuladas, como se observa en la Figura 4.5. Una vez dividida la lesión se cuantificaron el número de protuberancias y depresiones sustanciales, así como el índice de lobulación.

4.3.3 Mapa de distancias

El mapa de distancias puede ser usado para encontrar un máximo círculo inscrito que divida la lesión en áreas lobuladas, como se muestra en la Figura 4.6. Para ello se utilizó el algoritmo del punto medio [51], definiendo el radio y centro de la circunferencia mediante el valor máximo y su posición en el mapa de distancias, respectivamente.

Adicionalmente se utilizó el mapa de distancias para determinar parte de la masa exterior de la lesión así como el tejido circundante. Para ello se definió una distancia (k=3 [39]) hacia dentro y fuera de la lesión, respectivamente, como se observa en la Figura 4.7. A partir de esto se cuantificaron

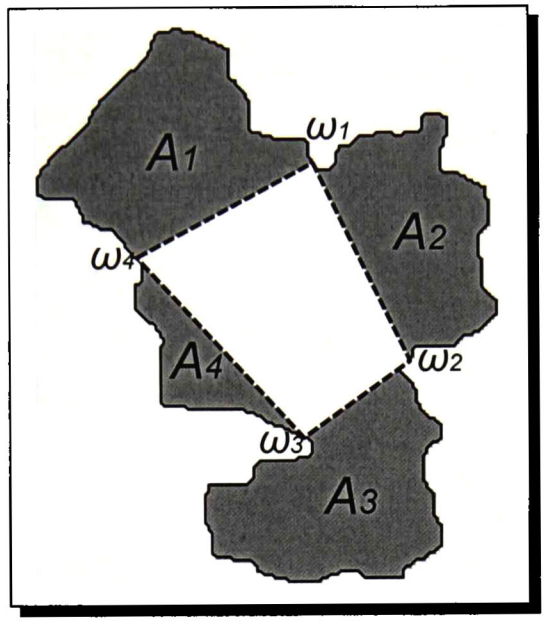


Figura 4.5: Lóbulos para un carcinoma, donde $\{\omega_1, \omega_2, \omega_3, \omega_4\}$ son cuatro puntos cóncavos y $\{A_1, A_2, A_3, A_4\}$ representan el área para cada lóbulo.

los parámetros: grado de interfaces abruptas, patrón del eco y patrón interno del eco, definidos en el Capítulo 3.

4.3.4 Elipse equivalente

Se implementó la elipse equivalente (EE) mediante la ecuación general de la elipse con centro en el origen y el cálculo de los momentos geométricos de la lesión, como se define en el Capítulo 3.

Una vez determinada la elipse equivalente, se calculó la disimilaridad para cada punto $P_1(x_1, y_1)$ en el borde de la lesión y su respectivo píxel de corte $P_2(x_2, y_2)$ en la EE, como se muestra en la Figura 4.8. A fin de determinar el punto $P_2(x_2, y_2)$ para cada punto $P_1(x_1, y_1)$, se implementó el Algoritmo 1.

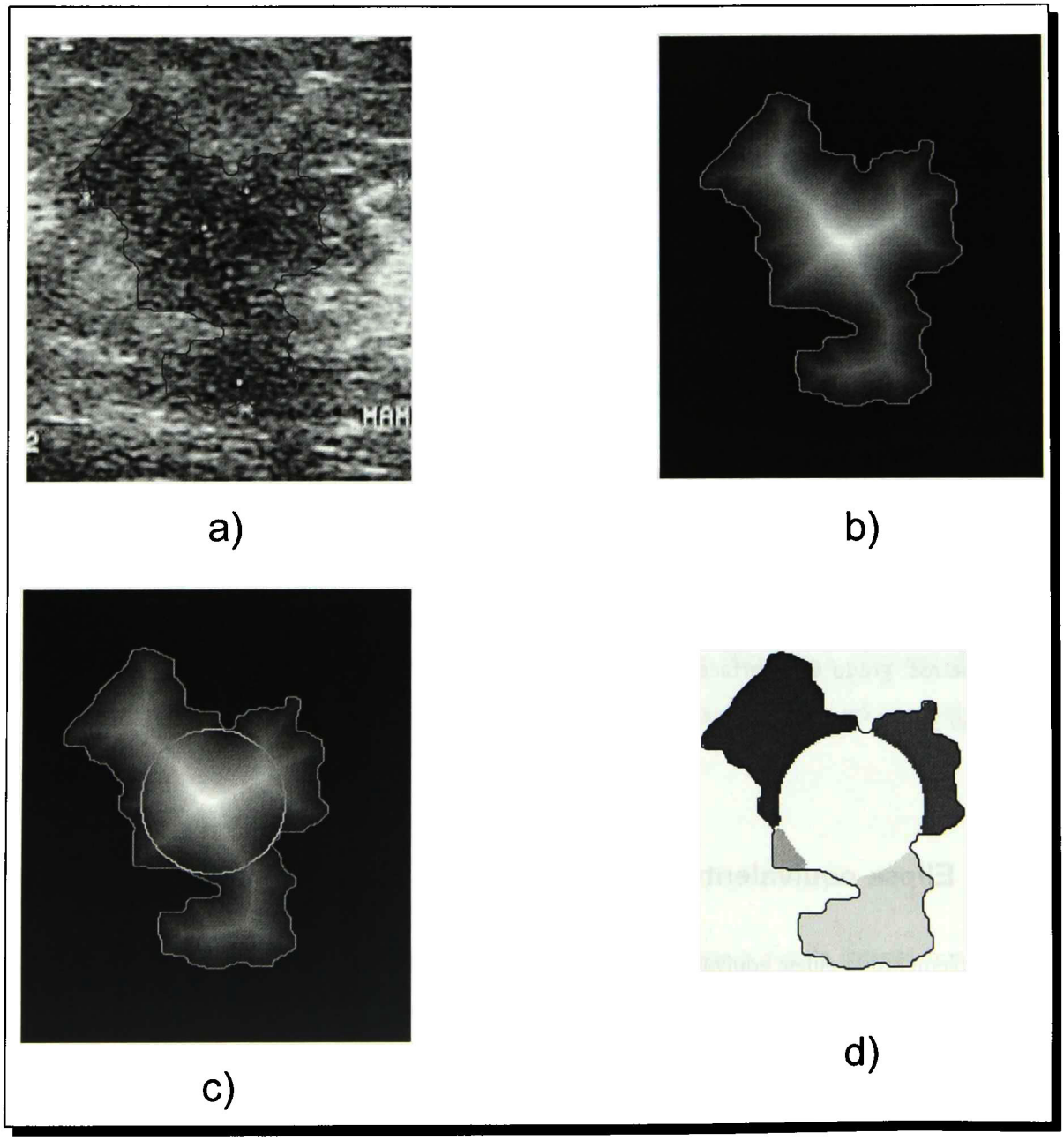


Figura 4.6: Ejemplo de (a) una lesión maligna, (b) mapa de distancias para (a), (c) máximo círculo inscrito determinado de (b) y (d) lobulaciones significantes.

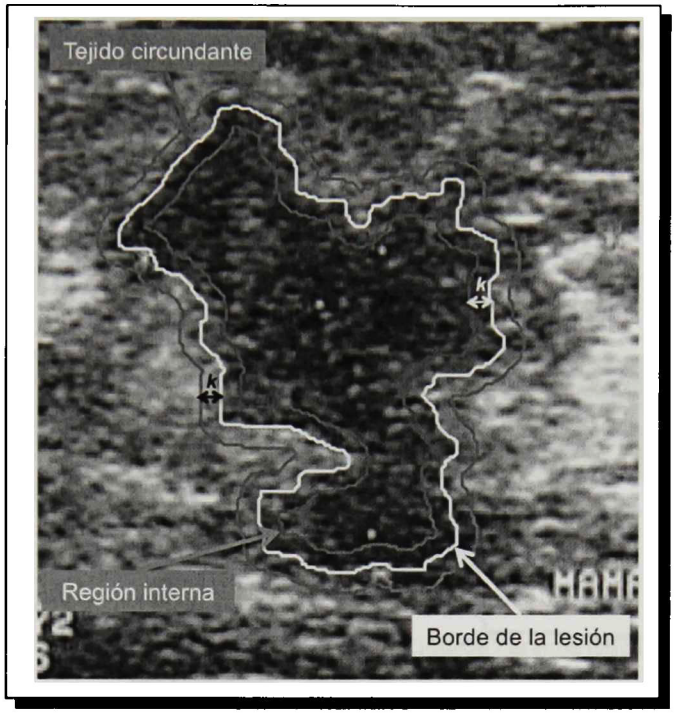


Figura 4.7: Masa exterior y tejido circundante para una lesión maligna.

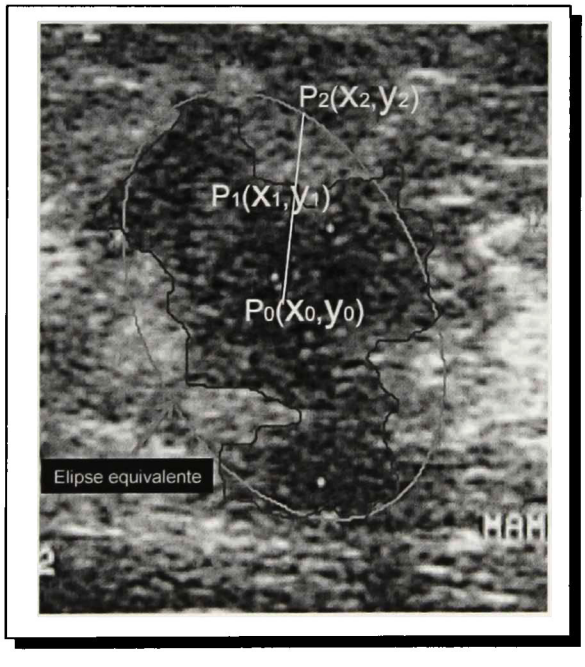


Figura 4.8: Disimilaridad entre contornos para un carcinoma y su elipse equivalente.

Algorithm 1 Punto de corte

Require: píxeles en el borde de la lesión (Bl), píxeles en la elipse equivalente (E)
Ensure: índices para cada punto de corte (ind).
1: $P_{Bl} \leftarrow$ Pendiente para cada elemento en Bl
2: $P_E \leftarrow$ Pendiente para cada elemento en E
3: $D \leftarrow$ Distancia entre P_{Bl} y P_E
4: $ind \leftarrow$ minimizar D
5: **return** ind

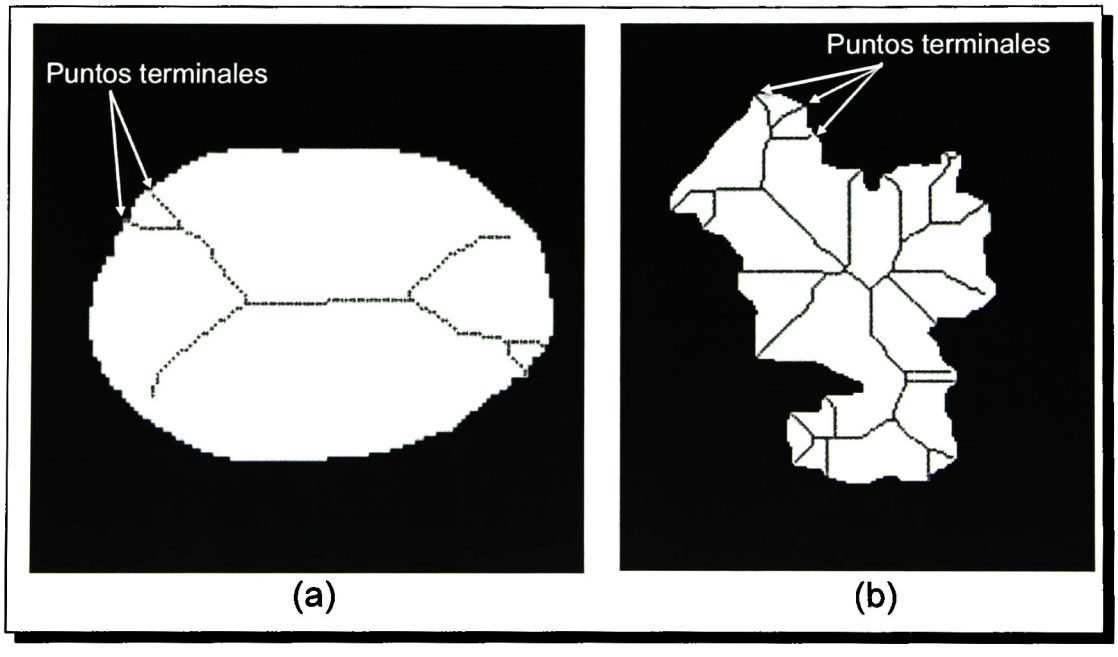


Figura 4.9: Ejemplo de esqueleto para (a)un carcinoma y (b) una lesión benigna.

4.3.5 Esqueleto

Se determinó el esqueleto de la lesión y se cuantificó la irregularidad de la misma mediante el número de puntos en el esqueleto. Adicionalmente se implementó un algoritmo basado en la operación *Hit-Miss* a fin de determinar el número de puntos terminales, como se observa en la Figura 4.9. Otro parámetro calculado fue el esqueleto elíptico normalizado, donde el factor de normalización es el numero de puntos en la elipse equivalente.

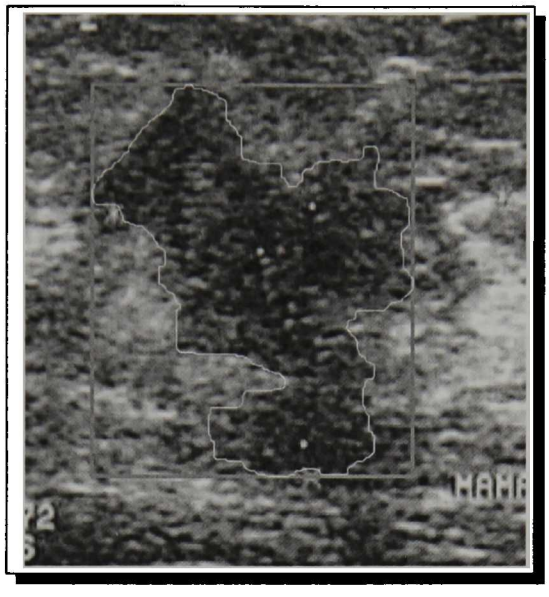


Figura 4.10: Región de interés para la extracción de características de textura.

4.3.6 Geométricos

Finalmente se calcularon los parámetros de circularidad, cociente morfológico, relación profundidad-ancho y tamaño de la lesión. Estos parámetros son comúnmente utilizados en la vida clínica diaria y están definidos en el Capítulo 3.

4.4 Extracción de características de textura

Los rasgos de textura fueron calculados de la imagen en escala de grises, donde para cada imagen se definió una RI (región de interés) mediante el cálculo de un mínimo rectángulo circunscrito que contiene a la lesión, como se muestra en la Figura 4.10.

Algorithm 2 Curva de complejidad

Require: Im (imagen de USM)
Ensure: C (curva de complejidad).

```
G ← Niveles de gris de la imagen
(X,Y) ← Dimensiones de la imagen
for g=0 to g=G-1 do
    Umb ← Im <g
    Umbx ← concatenar filas de Umb
    Umby ← concatenar columnas de Umb
    Tx ← número de transiciones en Umbx
    Ty ← número de transiciones en Umby
    C(g+1)=Tx+Ty;
end for
C=C/( X*(Y-1) + Y*(X-1) )
return C
```

4.4.1 Curva de complejidad

Se determinó la curva de complejidad mediante el Algoritmo 2. Como resultado se obtuvo una señal como la que se muestra en la Figura 4.11. Posteriormente se calculó el valor máximo de transiciones, valor medio de transiciones, media de la muestra, desviación estándar de la muestra y la entropía, a fin de describir el comportamiento de dicha señal. Estos estadísticos se definen en el Capítulo 3.

4.4.2 Matriz de co-ocurrencia de los niveles de gris

Una GLCM está definida por los parámetros distancia (d), orientación (θ) y el número de niveles de gris (L). En este trabajo se consideraron cinco distancias ($d = 1, 2, 3, 4, 5$), cuatro orientaciones ($\theta = 0°, 45°, 90°, 135°$) y una sola cuantificación de los niveles de gris ($L = 64$). Veinte GLCMs fueron construidas, a partir de las cuales se calcularon los 22 parámetros descritos en la Tabla 3.2, haciendo un total de 440 descriptores de textura extraídos de la GLCM.

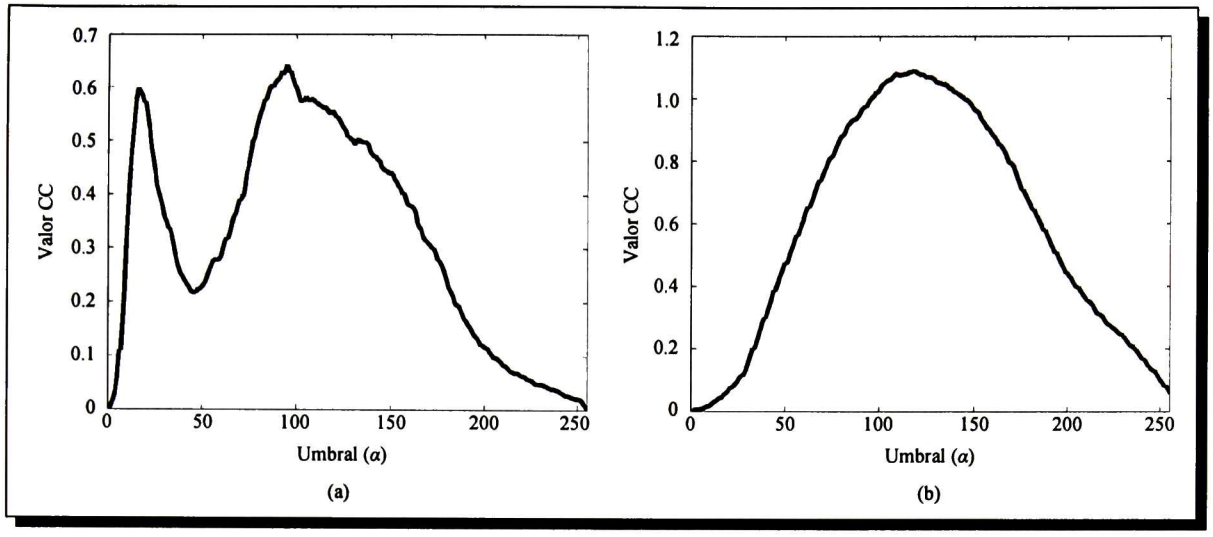


Figura 4.11: Curva de complejidad para (a) una lesión benigna y (b) un carcinoma.

4.4.3 Coeficientes normalizados de autocorrelación y autocovarianza

Para el cálculo de los coeficientes de autocorrelación y autocovarianza se utilizó una matriz de tamaño 5×5, obteniendo así 25 coeficientes. Posteriormente dichos coeficientes fueron normalizados utilizando como factor de normalización el primer coeficiente, el cual es posteriormente descartado ya que su valor siempre es la unidad, de modo que se obtuvieron 24 coeficientes de autocorrelación y 24 de autocovarianza.

4.4.4 Propiedades por bloques

La determinación de las características de textura por bloques se realizó mediante la división de la imagen en cuatro cuadrantes de igual tamaño, como se puede observar en la Figura 4.12. Una vez realizada la división de la imagen se calcularon los parámetros bloque de diferencia inversa de probabilidades y bloque de variación local de coeficientes de correlación, los cuales se definen en el Capítulo 3.

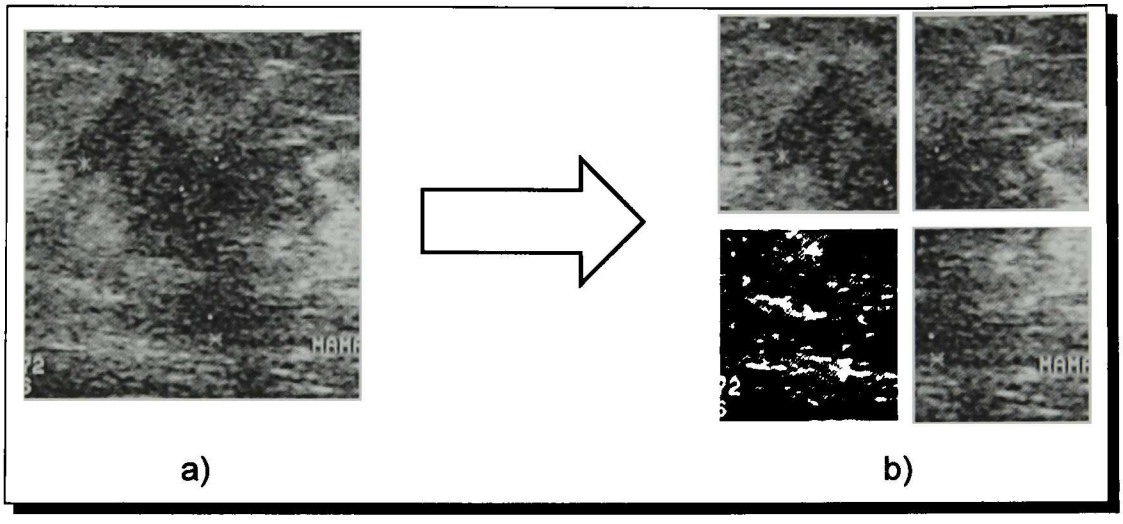


Figura 4.12: (a) Imagen de USM original, (b) división de la imágen para el cálculo de los descriptores basados en propiedades por bloques.

4.5 Normalización de los datos

Es necesario que los datos estén dentro de un determinado rango de valores para que los atributos sean comparables en algún sentido. Debido a que los parámetros calculados se encuentran en rangos distintos, es necesario realizar una transformación de los datos. Para ello se implementó la normalización **min-max**, la cual es una transformación lineal que escala el rango original de las variables a alguno especificado por el usuario como [52]:

$$y' = \left(\frac{y - \min_y}{\max_y - \min_y}\right) \cdot (\max_{y'} - \min_{y'}) + \min_{y'} \tag{4.3}$$

donde $\min_y$ y $\max_y$ son los valores mínimo y máximo, respectivamente para una muestra original y, y y' representa la muestra normalizada en un rango especificado por $\min_{y'}$ y $\max_{y'}$.

Para propósitos de la clasificación se estableció un rango de $[-1, 1]$ para cada característica calculada [53].

Índice	Identificador	Nombre
1	$DNRL$	Desviación estándar de la NRL
2	RA	Razón de área de la NRL
3	ENT	Entropía de la NRL
4	R	Rugosidad del contorno
5	RS	Razón de área de la envolvente convexa
6	NRV	Valor residual normalizado
7	$NSPD$	Número de protuberancias y lobulaciones sustanciales
8	LI	Índice de lobulación
9	MU	Número de lobulaciones significantes
10	LB_D	Grado de interfaces abruptas
11	EP_I	Patrón del eco
12	EP_{AG}	Patrón del eco interno
13	O_E	Orientación
14	SV	Disimilaridad
15	ENC	Anfractuosidad
16	$L:S$	Relación eje mayor:eje menor
17	$D:W$	Relación profundidad:ancho
18	C	Circularidad
19	$Mshape$	Cociente de cierre morfológico
20	SZ	Tamaño de la lesión
21	PTS	Número de puntos terminales
22	ENS	Esqueleto elíptico-normalizado

Tabla 4.1: Conjunto de descriptores morfológicos.

4.6 Construcción de espacios de características

Se construyeron tres espacios de características M-dimensionales: morfología ($M = 22$), textura ($M = 502$) y una combinación de ambos (M=524). En la Tabla 4.1 y la Tabla 4.2 se muestran los espacios de morfología y textura, respectivamente. Para el espacio combinado fueron concatenados el espacio morfológico y de textura.

Índice	Identificador	Nombre
1-20	$Autcr_{L,d,\theta}$	Autocorrelación
21-40	$Contr_{L,d,\theta}$	Contraste
41-60	$Corr1_{L,d,\theta}$	Correlación I
61-80	$Corr2_{L,d,\theta}$	Correlación II
81-100	$Protb_{L,d,\theta}$	Agrupamiento de protuberancias
101-120	$Sombr_{L,d,\theta}$	Agrupamiento de sombras
121-140	$Disim_{L,d,\theta}$	Disimilaridad
141-160	$Energ_{L,d,\theta}$	Energía
161-180	$Entrp_{L,d,\theta}$	Entropía
181-200	$Homo1_{L,d,\theta}$	Homogeneidad I
201-220	$Homo2_{L,D,\theta}$	Homogeneidad II
221-240	$DvStd_{L,d,\theta}$	Desviación estándar
241-260	$MxPrb_{L,d,\theta}$	Máxima probabilidad
261-280	$Scuad_{L,d,\theta}$	Suma de cuadrados
281-300	$SProm_{L,d,\theta}$	Suma de promedios
301-320	$SEntr_{L,d,\theta}$	Suma de entropía
321-340	$SVrnz_{L,d,\theta}$	Suma de varianza
341-360	$Dvrnz_{L,d,\theta}$	Diferencia de varianza
361-380	$DEntr_{L,d,\theta}$	Diferencia de entropía
381-400	$Info1_{L,d,\theta}$	Medida de información de correlación I
401-420	$Info2_{L,d,\theta}$	Medida de información correlación II
421-440	$DINor_{L,d,\theta}$	Diferencia inversa normalizada
441-464	$Acov_x$	Coeficientes de autocovarianza $(x = 1, 2 \ldots, 24)$
465-488	$Acov_x$	Coeficientes de autocovarianza $(x = 1, 2 \ldots, 24)$
489-492	$BVLC_x$	Bloque de variación local de coeficientes de correlación $(x = 1, 2, 3, 4)$
893-496	$BDIP_x$	Bloque de diferencia inversa de probabilidades $(x = 1, 2, 3, 4)$
497	$Vmax$	Valor máximo de transiciones
498	$Vmed$	Valor medio de transiciones
499	$Mmed$	Media de la muestra
500	$Mstd$	Desviación estándar de la muestra
501	Ent	Entropía
502	NRG	Gradiente radial normalizado

Tabla 4.2: Conjunto de descriptores de textura donde L representa el número de niveles de gris, d denota la distancia y θ la orientación para las GLCMs.

4.7 Selección de características

La selección de características no sólo implica la reducción de dimensionalidad, sino también la elección de atributos en función de su utilidad para el análisis. A fin de simplificar el proceso de selección de características es recomendable ordenar el espacio de características con base en su capacidad para describir el objeto, de este modo una vez ordenado se debe determinar cuántas de estas características serán consideradas en en el proceso de clasificación.

4.7.1 Ordenamiento de características

Dos de las principales técnicas de ordenamiento de características propuestas en la literatura para el problema de clasificación de lesiones de mama mediante ultrasonografía son el PCA y la técnica de MI.

4.7.1.1. Análisis de componentes principales

Es una transformación lineal que determina un nuevo sistema de coordenadas para el conjunto original de los datos, donde las varianzas mayores se encuentran en los primeros ejes, los que se denominan componentes principales. La implementación de esta técnica se realizó mediante el cálculo de la matriz de covarianza y su descomposición en autovalores como se muestra en el Algoritmo 3.

4.7.1.2. Información mutua

Esta técnica es comúnmente usada para el ordenamiento de espacios de características mediante el criterio de máxima-relevancia-mínima-redundancia. La condición de máxima relevancia busca seleccionar las características que comparten más información discriminante con la clase, mientras que la condición de mínima redundancia pretende reducir la redundancia entre las características seleccionadas. Esta implementación se realizó mediante la biblioteca desarrollada por Peng *et al.* [47].

Algorithm 3 Análisis de componentes principales

Require: Conjunto de datos $X = \{x_1, \ldots, x_M\}$ no vacío.
Ensure: Conjunto ordenado de datos $Y = \{y_1, \ldots, y_M\}$.
1: $X_{ajust} \leftarrow$ Datos con media cero
2: $C \leftarrow$ Matriz de covarianza de X_{ajust}
3: $AVec \leftarrow$ Auto-vectores de C
4: $AVal \leftarrow$ Auto-valores de C
5: $Vord \leftarrow$ Ordenar $AVec$ descendentemente mediante $AVal$
6: $Y \leftarrow X \times AVal$
7: **return** Y

Algorithm 4 Estimación de error

Require: Conjunto de datos ordenados $O = \{o_1, \ldots, o_M\}$
Ensure: Vector de errores para cada iteración Err.
1: $m \leftarrow 1$
2: **while** $m \leq M$ **do**
3: Seleccionar características de 1 a m
4: Crear 500 conjuntos *bootstrap*
5: Entrenar FLDA
6: Clasificar con FLDA.
7: $Err(m) \leftarrow$ Error *bootstrap*.632+
8: **end while**
9: **return** Err

4.7.2 Clasificación y estimación del error

Una vez ordenados cada uno de los tres conjuntos de características (morfología, textura y combinado) se realizó un clasificación iterativa (agregando características al proceso de clasificación). Se utilizó como clasificador el análisis lineal discriminante de Fisher debido a que no necesita de un proceso de sintonización de parámetros, además de ser un clasificador de fácil entrenamiento. Para determinar el desempeño de clasificación se utilizó el estimador *bootstrap* .632+ con 500 conjuntos para cada iteración. Este proceso se describe en el Algoritmo 4.

5

Resultados

5.1 Introducción

En este capítulo se presentan los resultados finales de la metodología propuesta en el Capítulo 4 para la selección de características en la clasificación de lesiones de mama. Adicionalmente se muestra una comparativa con los trabajos relacionados en el estado del arte.

5.2 Clasificación iterativa

En la Figura 5.1 se muestran los resultados para la clasificación iterativa de cada uno de los conjuntos evaluados. Los resultados revelan que el error de clasificación determinado por el estimador *bootstrap* .632+ tiende a disminuir en las primeras m características, alcanzando un valor mínimo, y aumenta conforme se agregan características irrelevantes o ruidosas. Además, se observa que con la técnica de MI se obtiene el menor error de clasificación con menos características que con la técnica de PCA.

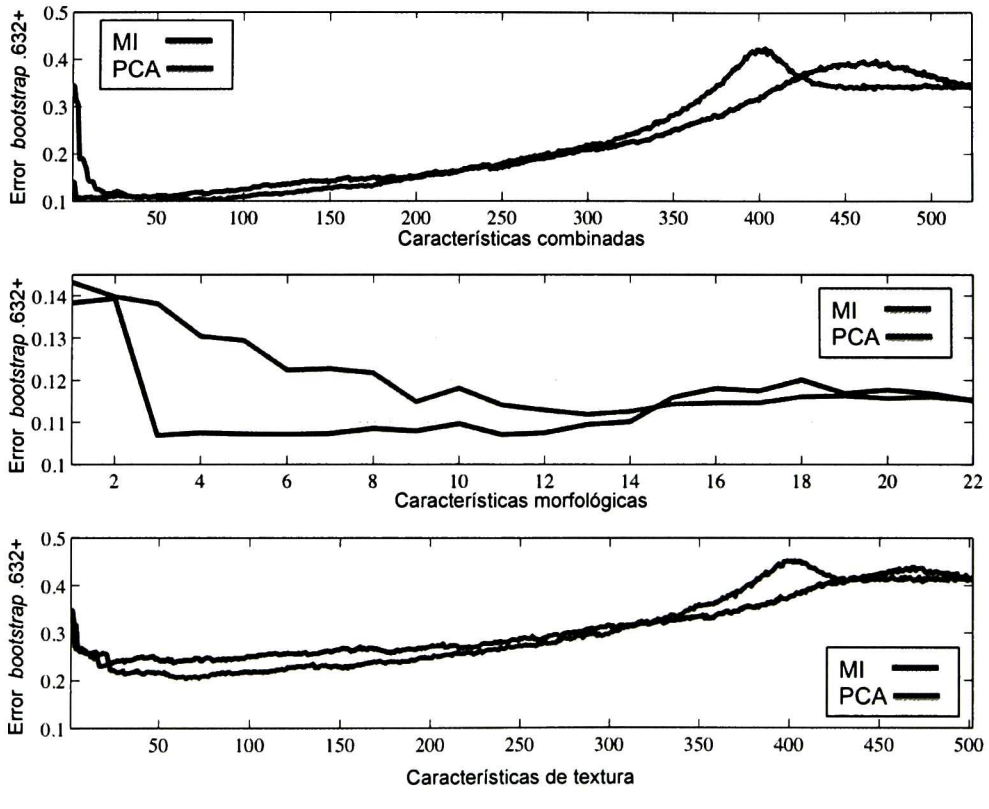


Figura 5.1: Error *bootstrap*.632+ para los conjuntos combinado, morfología y textura.

5.2.1 Determinación de los mejores subconjuntos

Una vez realizada la clasificación iterativa se determinó el número óptimo de características para cada subconjunto mediante la minimización del error *bootstrap*.632+ para cada una de las categorías con cada una de las técnicas de ordenamiento. En la Tabla 5.1 se muestra el error alcanzado con el número óptimo de atributos para cada una de las combinaciones entre las categorías de descriptores y las técnicas de ordenamiento.

	Combinado		Morfología		Textura	
	$M = 524$		$M = 22$		$M = 502$	
	PCA	MI	PCA	MI	PCA	MI
Error	0.100	0.153	0.111	0.106	0.204	0.236
m	69	13	13	3	65	24

Tabla 5.1: Número de características para el error mínimo en PCA y MI.

5.3 Métrica de desempeño

Para evaluar el desempeño de los subconjuntos seleccionados se empleó la métrica de área bajo la curva ROC (Az) que se define cómo la representación de la razón de verdaderos positivos (TPR, por su siglas en inglés) frente a la razón de falsos positivos (FPR, por su siglas en inglés). Dicha curva también se encuentra en función del umbral de discriminación (valor a partir del cual se decide que un caso es positivo). En el análisis ROC se definen cinco categorías que van desde un desempeño malo (Az=0.5) hasta un desempeño excelente (Az=1), como se muestra en la Figura 2.2.

5.4 Evaluación y resultados

La evaluación del desempeño se realizó mediante la creación de 500 nuevos conjuntos *bootstrap* utilizando el FLDA como clasificador. Se utilizó la prueba de Shapiro-Wilk (α=0.05) [54] [55] a fin de determinar si los resultados obtenidos tienen una distribución normal. El resultado de esta prueba determinó que los datos no son normales, es decir, que su distribución es asimétrica, por lo que se optó por utilizar estadísticos robustos como la mediana (MD), para estimar tendencia central, y el estimador Qn, para cuantificar la dispersión [56].

Se comparó el poder de discriminación del conjunto de características completo (M atributos) y los subconjuntos determinados por PCA y MI (m atributos) para cada espacio de características mediante la prueba estadística no paramétrica de Kruskal-Wallis ($\alpha = 0.05$) [57] y el análisis multicomparativo por pares mediante el método de Bonferroni, como se observa en la Figura 5.2, donde el asterisco '*' indica que las medianas de los grupos no son significativamente diferentes.

Los resultados de la prueba estadística se muestran en la Tabla 5.2 en términos de MD y Qn, de donde se observa lo siguiente:

1. Para las características de textura se mejoró la mediana de Az de 0.588 para el conjunto completo (M=502) a 0.840 para PCA (m=65) y 0.820 para MI (m=24).

2. Para el conjunto morfológico se mantuvo el desempeño Az de 0.948 para el conjunto completo (M=22), 0.946 para PCA (m=13) y 0.943 para MI (m=3).

Finalmente se determinó que el mejor subconjunto de características fue el subconjunto combinado con la técnica de MI con un desempeño de Az=0.95 y sólo 13 características. Sin embargo, tanto el conjunto morfológico completo (22 características) como el subconjunto morfológico arrojado

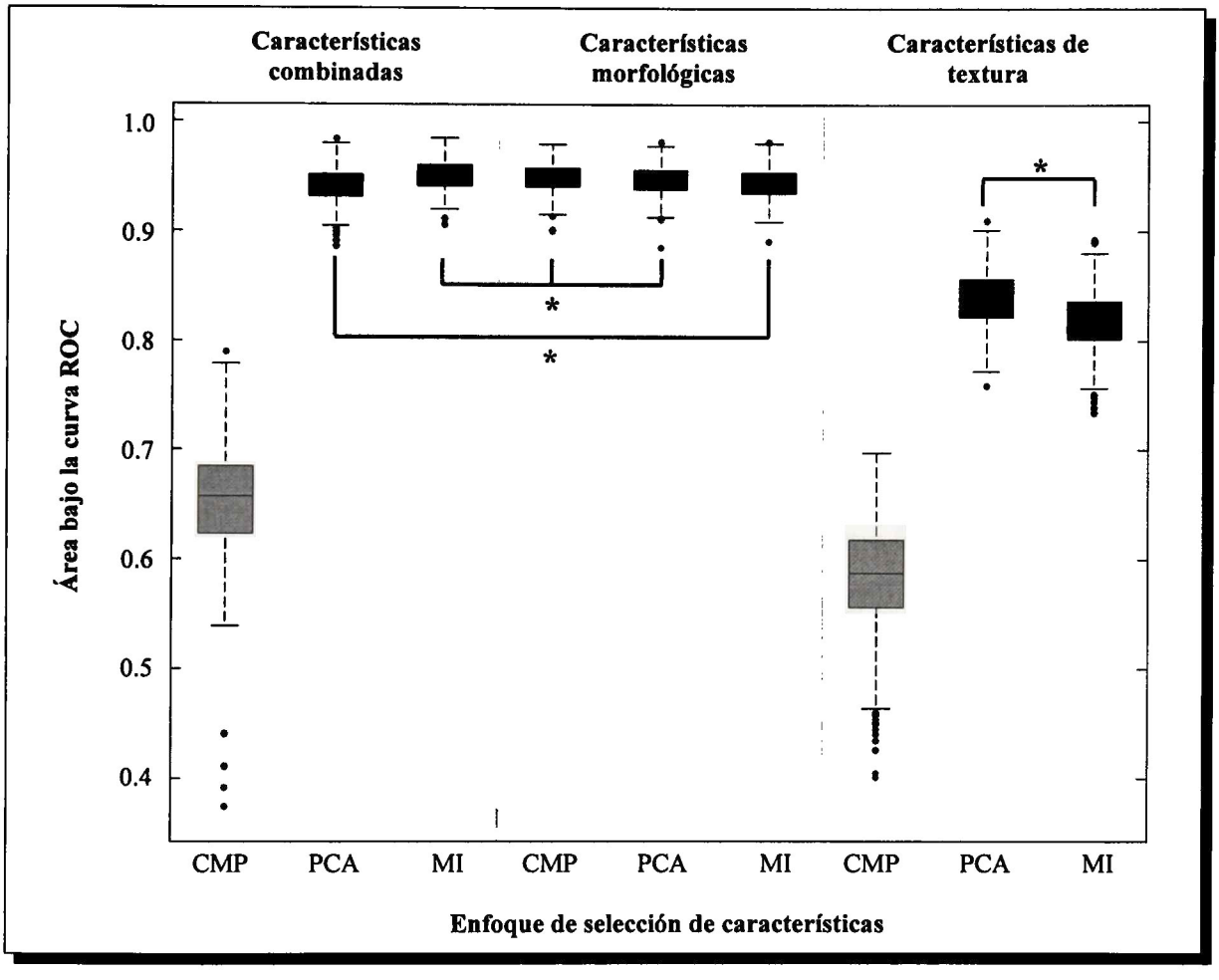


Figura 5.2: Distribución de valores de área bajo la curva para los conjuntos completos (CMP) y los subconjuntos seleccionados considerando PCA y MI.

Combinado			
	Completo	PCA^b	MI^a
MD	0.657	0.941	0.951
Q_n	0.045	0.014	0.013
Morfología			
	Completo^a	PCA^a	MI^b
MD	0.948	0.946	0.943
Q_n	0.012	0.013	0.013
Textura			
	Completo	PCA^c	MI^c
MD	0.588	0.840	0.820
Q_n	0.048	0.026	0.025

Tabla 5.2: Área bajo la curva (MD y Qn) para cada uno de los grupos evaluados, donde el mismo superíndice indica los grupos que no son significativamente diferentes.

por la técnica de PCA (13 características) obtuvieron desempeños de clasificación estadísticamente similares.

En la Tabla 5.3 se listan los descriptores seleccionados por el método propuesto, así como el tipo de descriptor (morfología o textura) y la técnica de la cual provienen.

5.5 Comparativa

Se implementó el método propuesto en el artículo "*Complexity curve and grey level co-occurrence matrix in the texture evaluation of breast tumor on ultrasound images*" propuesto por Alvarenga *et al.* [14] debido a que se cuenta con el método de segmentación propuesto en el artículo, además de usar el clasificador FLDA, usado también en esta tesis. En dicho enfoque los autores proponen la extracción de características de textura a partir de la GLCM y la CC. Para ello definen dos regiones para cada imagen (región de interés y región interna del tumor). Cinco parámetros fueron obtenidos de la GLCM y otros cinco de la CC para cada región, haciendo un total de 20 descriptores.

Descriptor	Tipo	Categoría
Esqueleto elíptico-normalizado	M	Esqueleto
Número de protuberancias y depresiones sustanciales	M	Envolvente convexa
Razón profundidad:ancho	M	Geométricos
Razón de área de la envolvente convexa	M	Envolvente convexa
Orientación	M	Elipse equivalente
Diferencia de entropía $(L = 64, d = 4, \theta = 45)$	T	Matriz de co-ocurrencia
Relación eje mayor:eje menor	M	Elipse equivalente
Auto-covarianza (1,2)	T	Coeficientes de auto-covarianza
Auto-covarianza (0,1)	T	Coeficientes de auto-covarianza
Auto-covarianza (4,2)	T	Coeficientes de auto-covarianza
Media de la muestra	T	Curva de complejidad
Variación de coeficientes de correlación	T	Propiedades por bloques
Energía $(L = 64, d = 4, \theta = 45)$	T	Matriz de co-ocurrencia

Tabla 5.3: Mejor subconjunto de características encontrado.

Adicionalmente se aplicó la metodología de selección de características propuesta en esta tesis, con el objetivo de determinar si el error de clasificación disminuía. La evaluación se realizó en términos del error *bootstrap .632+* utilizando las 638 ultrasonografías de nuestro banco de imágenes. En la Tabla 5.4 se muestran los resultados obtenidos.

Referencia	Número de características	Error .632+
Alvarenga *et al.* 2007 (completo)	20	0.2132
Alvarenga *et al.* 2007 (mejor obtenido)	5	0.2022
Método propuesto (PCA)	19	0.1994
Método propuesto (MI)	2	0.1986

Tabla 5.4: Resultados para la implementación de un enfoque propuesto en la literatura

Como se puede observar en la Tabla 5.4, se superó el desempeño de la implementación realizada reduciendo el espacio de características así como el error *bootstrap* .632+ obtenido. El mejor subconjunto de características propuesto por Alvarenga *et al.* [14] posee cinco atributos con un error *Bootstrap* .632+ de 0.20, mientras que con la técnica de MI el número de características se redujo a dos atributos con un error de 0.19.

Autores	NI	NC	TD	MS	Desempeño de clasificación
Huang *et al.* [20]	118 (84-B, 34-M)	19	morfológicos	PCA	Az=0.89
Gómez *et al.* [23]	641 (413-B, 228-M)	22	morfológico	MI	Az=0.87
Pereira *et al.* [22]	246 (69-B, 177-M)	7	morfológicos	MI	Az=0.86
Wan *et al.* [31]	321 (113-B, 208-M)	1285	combinación	PCA	Az=0.82
				RPCA	Az=0.92
Gómez *et al.* [32]	436 (219-B, 217-M)	17	textura	mRMR	Az=0.87
Rivera y Gómez [33]	960	125	textura	MI(mRMR)	Az=0.81
Wu *et al.* [34]	210 (120-B, 90-M)	5	combinación	GA	Az=0.96
Shen *et al.* [58]	168 (104-B, 64-M)	23	morfológico	MI(mRMR)	Az=0.85
Método propuesto	638(413-B, 225-M)	13	combinación	MI(mRMR)	Az=0.95

Tabla 5.5: Comparativa del método propuesto con el estado del arte. NI: Número de imágenes, NC: Número de características, TD: Tipo de descriptor, MS: Método de selección.

Por último, de acuerdo a los resultados presentados en el estado del arte (para los trabajos con un enfoque de selección de características), en términos de Az, en general se superó el desempeño de clasificación para la mayoría de los enfoques mediante la metodología propuesta, como se muestra en la Tabla 5.5.

6

Conclusiones y trabajo futuro

6.1 Conclusiones

Actualmente el cáncer de mama es una de las principales causas de muerte en mujeres de todo el mundo. Debido a que las causas del cáncer de mama aún se mantienen desconocidas la Organización Mundial de la Salud recomienda a los gobiernos de cada país implementar estrategias de diagnóstico oportuno, dentro de estas estrategias destacan la mamografía como principal fuente de diagnóstico y el USM como técnica coadyuvante a la mamografía. El objetivo principal del diagnóstico por USM es mejorar la interpretación del diagnóstico en general por parte de los radiólogos. El objetivo es evitar biopsias innecesarias en lesiones benignas evitando así la presión psicológica producida al paciente debido a que es un procedimiento invasivo, así como reducir los costos económicos producto de estos estudios. Para ello se han propuesto sistemas de diagnóstico asistido por computadora, los cuales se basan en cuatro etapas básicas: 1) segmentación, 2) preprocesamiento, 3) extracción y selección de características y 4) clasificación.

Técnica	Tipo	No. descriptores
Longitud radial normalizada	Morfología	4
Envolvente convexa	Morfología	4
Razón Elipse equivalente	Morfología	4
mapa de distancias	Morfología	4
Esqueleto	Morfología	2
Geométricos	Morfología	4
Curva de complejidad	Textura	5
Coeficientes de auto-correlación	Textura	24
Coeficientes de auto-covarianza	Textura	24
Propiedades por bloques	Textura	8
Matriz de co-ocurrencia de los niveles de gris	Textura	440
Gradiente radial Normalizado	Textura	1

Tabla 6.1: Técnicas de descripción para lesiones de USM.

En esta tesis se presenta una metodología para la extracción y selección de características en la clasificación de lesiones de USM con el objetivo de demostrar que para un conjunto completo M-dimensional de características morfológicas y de textura existe un subconjunto m-dimensional de características (donde $m < M$) que mejore o mantenga el desempeño de clasificación.

Se calcularon 22 descriptores morfológicos y 502 de textura mediante la implementación de diferentes técnicas de descripción. En la Tabla 6.1 se hace un resumen del número de atributos extraídos para cada técnica. Asimismo, se utilizó un banco de imágenes con 638 ultrasonografías (413 benignas y 225 malignas), las cuales fueron diagnosticadas mediante biopsia.

Tres espacios de características M-dimensionales fueron creados: morfología (M=22), textura (M=502) y una combinación de ambos (M=524). Cada uno de estos conjuntos fueron normalizados en el rango [-1,1] y posteriormente ordenados mediante las técnicas de PCA y MI, donde las primera características en el espacio ordenado presentan un mayor poder de discriminación entre las clases

de lesión benigna y maligna.

Para cada espacio M-dimensional se crearon M subconjuntos agregando iterativamente las primeras m características al proceso de clasificación hasta considerar el conjunto completo de atributos. Se evaluó el desempeño de un clasificador FLDA mediante el estimador *bootstrap*.632+ utilizando 500 grupos (entrenamiento y prueba) para cada subconjunto de características creado.

Los resultados revelan que el error de clasificación determinado por el estimador *bootstrap*.632+, tiende a disminuir en las primeras m características, alcanzando un valor mínimo, y aumenta conforme se agregan características irrelevantes o ruidosas. Por tanto, el mejor subconjunto de características se define como aquel donde se minimiza la curva de error.

Una vez determinado el mejor subconjunto de características para cada espacio probado se evaluó el desempeño de cada uno de ellos mediante la métrica de área bajo la curva (Az). Posteriormente se realizó la prueba de Shapiro-Wilk ($\alpha = 0.05$) para determinar la normalidad de los datos. Se observó que algunos grupos presentaron distribución asimétrica, por lo que se optó por utilizar estadísticos robustos como la mediana (MD) y el estimador Qn.

Se comparó el poder de discriminación del conjunto de características completo (M atributos) y los subconjuntos determinados por PCA y MI (m atributos) mediante la prueba estadística de Kruskal-Wallis ($\alpha = 0.05$). Los resultados apuntan que, seleccionando un subconjunto de características adecuado, es posible mejorar o mantener el desempeño de clasificación. Para las características de textura se logró eliminar características ruidosas, mejorando la mediana de Az de 0.588 para el conjunto completo (M=502) a 0.840 para PCA (m=65) y 0.820 para MI (m=24). Por otro lado, para el conjunto morfológico se mantuvo el desempeño de 0.948 para el conjunto completo (M=22) a 0.946 para PCA (m=13) y 0.943 para MI (m=3) mediante la eliminación de características

redundantes.

El mejor valor de mediana para área bajo la curva se alcanzó mediante el subconjunto de descriptores combinados ($m = 13$) obtenido por la técnica de MI. Esto representa el 2.5 % del espacio de características completo (M=524). Adicionalmente, tanto el conjunto morfológico completo (M=22) como el subconjunto morfológico obtenido por la técnica de PCA (m=13) obtuvieron desempeños de clasificación estadísticamente similares que el mejor subconjunto obtenido.

Finalmente se cree que las características de textura aportan menor información discriminante frente a las morfológicas debido a que estas son calculadas sobre regiones muy pequeñas de la imagen, sin tomar en cuenta las macrotexturas, tal como lo hace un radiólogo en su diagnóstico; mientras que las características morfológicas se calculan sobre toda la región de la lesión.

6.2 Trabajo futuro

Si bien es cierto que el mejor subconjunto de características seleccionado superó el desempeño de clasificación de otros enfoques de selección de características propuestas en la literatura, se cree que es posible incrementar el desempeño mediante el uso de otro tipo de parámetros como los descriptores basados en modelos, además del uso de otras técnicas de selección de características tales como los algoritmos genéticos.

Bibliografía

[1] OMS, "Organización mundial de la salud." http://www.who.int, Septiembre 2013. 1

[2] INEGI, "Instituto nacional de estadística y geografía," in *http://www.inegi.org.mx/inegi/contenidos/espanol/prensa/contenidos/estadisticas/2011/cancer1*..... 2011. 2

[3] M. E. Brandan and Y. V. Navarro, "Detección del cáncer de mama: Estado de la mamografía en méxico," *Cancerología*, vol. 1, pp. 147–162, 2006. 2

[4] H. Cheng, J. Shan, W. Ju, Y. Guo, and L. Zhang, "Automated breast cancer detection and classification using ultrasound images: a survey," *Pattern Recognition*, vol. 43, pp. 299–317, 2010. 2, 3

[5] S. AT, T. D, R. CL, D. MA, P. SH, and S. GA, "Solid breast nodules: use of sonography to distinguish between benign and malignant lesions," *Radiology*, vol. 196, no. 1, pp. 123–134, 1995. 3

[6] J. Timmers, H. van Doorne-Nagtegaal, A. Verbeek, G. den Heeten, and M. Broeders, "A dedicated bi-rads training programme: Effect on the inter-observer variation among screening radiologists," *European Journal of Radiology*, vol. 81, no. 9, pp. 2184–2188, 2012. 3

[7] C. MJ, A. RM, G. B, and P. WC., "Intraobserver interpretation of breast ultrasonography following the bi-rads classification," *European Journal of Radiology*, vol. 74, no. 3, pp. 525–528, 2010. 3

[8] C. SC, C. YC, C. M. Su CH, H. TL, and H. S, "Analysis of sonographic features for the

differentiation of benign and malignant breast tumors of different sizes," *journal*, vol. 23, no. 2, pp. 188–193, 2004. 4

[9] Y. Yue, M. M. Croitoru, A. Bidani, J. B. Zwischenberger, and J. W. Clark, "Nonlinear multiscale wavelet diffusion for speckle suppression and edge enhancement in ultrasound images," *IEEE Transaction on Medical Imaging*, vol. 25, pp. 297–311, 2006. 4

[10] P. Berkhin, "Survey of clustering data mining techniques," tech. rep., Accrue Software, San Jose, CA, 2002. 5

[11] W. Gómez, L. Leija, A. V. Alvarenga, A. F. C. Infantosi, and W. C. A. Pereira, "Computerized lesion segmentation of breast ultrasound based on marker-controlled watershed transformation," *Medical Physics*, vol. 37, pp. 82–95, January 2010. 7, 53

[12] H. Chenga, J. Shan, W. Ju, Y. Guo, and L. Zhang, "Automated breast cancer detection and classification using ultrasound images: A survey," *Pattern Recognition*, vol. 43, pp. 299–317, 2010. 11

[13] R.-F. Chang, W.-J. Wu, W. K. Moon, and D.-R. Chen, "Improvement in breast tumor discrimination by support vector machines and speckle-emphasis texture analysis," *Ultrasound in Medicine and Biology*, vol. 29, no. 5, pp. 679–686, 2003. 12, 24, 39

[14] A. V. Alvarenga, W. C. A. Pereira, and A. F. C. Infantosi, "Complexity curve and grey level co-occurrence matrix in the texture evaluation of breast tumor on ultrasound images," *Medical Physics*, vol. 34, pp. 379–387, 2007. 13, 24, 35, 36, 37, 74, 75

[15] Y.-Y. Liao, J.-C. Wu, C.-H. Li, and C.-K. Yeh, "texture feature analysis for breast ultrasound image enhancement," *Ultrasonic Imaging*, vol. 33, no. 4, pp. 264–278, 2011. 13, 24

[16] B. Liu, H.D.Cheng, JianhuaHuang, JiaweiTian, X. Tang, and JiafengLiu, "Fully automatic and

segmentation-robust classification of breast tumors based on local texture analysis of ultrasound images," *Pattern Recognition*, vol. 43, pp. 280–298, 2010. 13, 24

[17] R.-F. Chang, W.-J. Wu, W. K. Moon, and D.-R. Chen, "Automatic ultrasound segmentation and morphology based diagnosis of solid breast tumors," *Breast Cancer Research and Treatment*, vol. 89, no. 2, pp. 179–185, 2005. 14, 24

[18] R. N. Czerwinski, D. L. Jones, and W. D. O'Brien, "Detection of lines and boundaries in speckle images—application to medical ultrasound," *IEEE Transaction on Medical Imaging*, vol. 18, pp. 126–136, February 199. 14

[19] O. N, "A threshold selection method form gray-level histograms," *IEEE Trans Systems, Man, and Cybernetics*, vol. 39, no. 3, pp. 62–66, 1979. 14

[20] Y.-L. Huang, D.-R. Chen, Y.-R. Jiang, S.-J. Kuo, H.-K. Wu, and W. K. Moon, "Computer-aided diagnosis using morphological features for classifying breast lesions on ultrasound," *Ultrasound Obstet Gynecol*, vol. 32, pp. 565–572, 2008. 14, 24, 76

[21] H. Y. . J. Y. . C. D. . M. WK, "Level set contouring for breast tumor in sonography," *IEEE Trans Systems, Man, and Cybernetics*, vol. 20, no. 3, pp. 238–247, 2007. 14

[22] A. V. Alvarenga, A. F. C. Infantosi, W. C. A. Pereira, and C. M. Azevedo, "Assessing the performance of morphological parameters in distinguishing breast tumors on ultrasound images," *Medical Engineering & Physics*, vol. 43, pp. 49–56, 2010. 15, 24, 26, 27, 28, 34, 76

[23] W. Gómez, W. C. A. Pereira, A. F. C. Infantosi, and A. D. Pérez, "Computerized diagnosis of breast lesions on ultrasonography," *XXII Brazilian Congress on Biomedical Engineering, CBEB 2010*, pp. 399–402, 2010. 16, 76

[24] L. Bocchi, F. Gritti, C. Manfredi, E. Giannotti, and J. Nori, "Semiautomated breast cancer classification from ultrasound video," *IEEE*, vol. volume, pp. 1112–1115, 2012. 16, 24

[25] W.-J. Wu and W. K. Moon, "Ultrasound breast tumor image computer-aided diagnosis with texture and morphological features1," *Academic Radiology*, vol. 15, pp. 873–880, 2008. 17, 24

[26] Y.-Y. Liao and C.-K. Yeh, "An integrated approach based on morphology, texture, and backscattering-statistics for distinguishing between benign and malignant breast," *Ultrasonics Symposium (IUS), 2010 IEEE*, pp. 1408–1411, 2010. 17, 24

[27] L. Yan, C. Heng-Da, Huang, Jianhua, Z. Yingtao, T. Xianglong, W. Hong, and T. Jiawei, "Computer-aided diagnosis system for breast cancer using b-mode and color doppler flow images," *Optical Engineering*, vol. 51, no. 4, pp. 043202–043202–9, 2012. 17, 24

[28] W. K. Moon, C. M. Lo, J. M. Chang, C.-S. Huang, J.-H. Chen, and R.-F. Chang, "Computer-aided classification of breast masses using speckle features of automated breast ultrasound images," *Med. Phys*, vol. 39, no. 10, pp. 6465–6473, 2012. 18, 24

[29] W. CoelhoA.Pereira, A. V. Alvarenga, A. C.Infantosi, L. Macrini, and C. E.Pedreira, "A non-linearmorphometric feature selection approach for breast tumor contour from ultrasonic images," *Computers in Biology and Medicine*, vol. 40, pp. 912–918, 2010. 18, 24

[30] J. Wright, Y. Peng, Y. Ma, A. Ganesh, and S. Rao, "Robust principal component analysis: Exact recovery of corrupted low-rank matrices by convex optimization," in *Advances in Neural Info. Process. Systems*, 2009. 18

[31] T. Wan, R. Liao, and Z. Qin, "A robust feature selection approach using low rank matrices for breast tumors in ultrasonic images," *IEEE International Conference on Image Processing*, vol. volume, pp. 1645–1648, 2011. 18, 24, 76

[32] W. Gómez, W. C. A. Pereira, and A. F. C. Infantosi, "Analysis of co-occurrence texture statistics as a function of gray-level quantization for classifying breast ultrasound," *IEEE Transaction on Medical Imaging*, vol. 31, pp. 1889–1899, 2012. 19, 24, 76

[33] I. Rivera-Islas and W. Gómez, "Analytical study of texture features based on gray-level co-ocurrence matrix for automatic segmentation of breast ultrasound," *XXIII Brazilian Congress on Biomedical Engineering*, 2012. 19, 24, 76

[34] W.-J. Wu, A.-W. Lin, and W. K. Moon, "Combining support vector machine with genetic algorithm to classify ultrasound breast tumor images," *Computerized Medical Imaging and Graphics*, p. in press, 2012. 19, 24, 76

[35] U. SCHEIPERS, C. PERREY, C. H. STEFAN SIEBERS, and H. ERMERT, "A tutorial on the use of roc analysis for computer-aided diagnostic systems," *ULTRASONIC IMAGING*, 2005. 21

[36] Y. Wang, J. Shen, Yi, Guo, and W. Wang, "Computerized classification of breast tumors with morphologic and texture features of ultrasonic images," *IEEE International Symposium on Computer-Based Medical Systems*, vol. volume, pp. 23–28, 2008. 24

[37] H. Chiang, C.-M. Tiu, G.-S. Hung, S.-C. Wu, T. Chang, and Y.-H. Chou, "Stepwise logistic regression analysis of tumor contour features for breast ultrasound diagnosis," in *Ultrasonics Symposium, 2001 IEEE*, vol. 2, pp. 1303–1306 vol.2, 2001. 26, 27, 34

[38] C.-M. C. . Y.-H. C. . K.-C. H. . G.-S. H. . C.-M. T. . H.-J. C. . S.-Y. Chiou, "Breast lesions on sonograms: Computer-aided diagnosis with nearly setting- independent features and artificial neural networks," *Radiology*, 2003. 28, 29, 31, 33, 34

[39] W.-C. Shen, R.-F. Chang, W. K. Moon, Y.-H. Chou, and C.-S. Huang, "Breast ultrasound computer-aided diagnosis using bi-rads features," *Academic Radiology*, vol. 14, pp. 928–939, August 2007. 30, 31, 32, 33, 56

[40] S. Patnaik and Y.-M. Yang, eds., *Soft Computing Techniques in Vision Science*, vol. 395. Springer, 2012. 33

[41] L.-K. Soh and C. Tsatsoulis, "Texture analysis of sar sea ice imagery using gray level co-ocurrence matrices," *IEEE transactions on Geoscience and Remote Sensing*, 199. 36, 37

[42] M. Bevk and I. Kononenko, "A statistical approach to texture description of medical images: A preliminary study," *IEEE symposium on Computer-Based Medical Systems*, 2002. 36

[43] R. M. Haralick, K. Shanmugam, and Its'HakvDisntein, "Textural features for image classification," *IEEE transactions on Systems, Man and Cybernetics*, vol. 3, no. 6, pp. 610–621, 1973. 37

[44] D. A. Clausi, "An analysis of co-occurrence texture statistics as a function of grey level quantization," *Can. J. Remote Sensing*, vol. 28, no. 1, pp. 45–62, 2002. 37

[45] Y. H. . K.-L. W. . D.-R. Chen, "Diagnosis of breast tumors with ultrasonic texture analysis using support vector machines," *Neural Comput Applic*, vol. 15, no. 3, pp. 164–169, 2005. 39, 41

[46] J. Shlens, "A tutorial on principal component analysis," tech. rep., Center for Neural Science, New York University, 2009. 42

[47] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, pp. 1226–1238, August 2005. 43, 67

[48] G. Junying and Z. Youwei, "Generalized kernel function fisher discriminant for pattern recognition," in *6th International Conference on Signal Processing*, vol. 2, pp. 1075–1078, 2002. 45

[49] B. Efron and R. Tibshirani, "Improvements on cross validation: The .632+ bootstrap method," *American Statistical Association*, vol. 92, pp. 548–560, June 1997. 46, 48

[50] A. M. Molinaro, R. Simon, and R. M. Pfeiffer, "Prediction error estimation: a comparison of resampling methods," *Bioinformatics*, vol. 21, no. 15, pp. 3301–3307, 2005. 48

[51] J. R. V. Aken, "An efficient ellipse-drawing algorithm," *IEE CG&A*, pp. 24–35, September 1984. 56

[52] W. li and Z. Liu, "A method of svm with normalization in intrusion detection," *Procedia Environmental Sciences*, vol. 11, pp. 256–262, 2011. 64

[53] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*. Academic Press, 4th ed., 2009. 64

[54] S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normallity (complete samples)," *Biometrika*, vol. 52, pp. 591–611, December 1965. 72

[55] N. M. Razali and Y. B. Wah, "Power comparisons of shapiro-wilk, kolmogorov-smirnov, liliefors and anderson-darling tests," *Statistical Modeling and Analytics*, vol. 2, no. 1, pp. 21–33, 2011. 72

[56] P. J. Rousseeuw and C. Croux, "Alternatives to the median absolute deviation," *American Statistical Association*, vol. 88, pp. 1273–1283, December 1993. 72

[57] W. H. Kruskal and W. A. Wallis, "Use of ranks in one-criterion variance analysis," *American Statistical Association*, vol. 47, pp. 583–621, December 1952. 72

[58] X. Shen, S. Zhang, R. Yao, Y. Chen, Y.-M. Zhu, and S. Zhang, "Discrimination between benign and malignant breast cancers in ultrasound images based on cost-sensitive boosting," *Intelligent science and intelligent data engineering*, vol. 7202, pp. 136–144, 2012. 76

CENTRO DE INVESTIGACIÓN Y DE ESTUDIOS AVANZADOS DEL IPN

UNIDAD TAMAULIPAS

Cd. Victoria, Tamaulipas, a 7 de febrero de 2014.

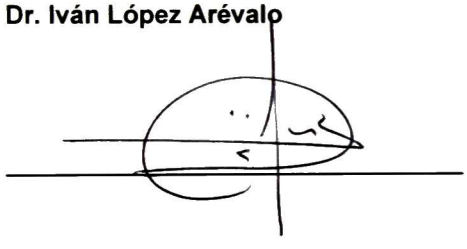
Los abajo firmantes, integrantes del jurado para el examen de grado que sustentará el C. Cristhian Muñoz Meza, declaramos que hemos revisado la tesis titulada:

"Estudio comparativo de técnicas de selección de características para la clasificación de lesiones de mama en ultrasonografía"

Y consideramos que cumple con los requisitos para obtener el grado de Maestro en Ciencias en Computación.

Atentamente,

Dr. Iván López Arévalo



Dr. César Torres Huitzil

Dr. Wilfrido Gómez Flores

Km. 5.5 Carretera a Soto la Marina. Ciudad Victoria, Tamaulipas. CP 87130
Parque Científico y Tecnológico TECNOTAM
Tels.: (834) 1070220 y 316 6600
E-mail: admin@tamps.cinvestav.mx