UNIVERSIDAD TECNOLÓGICA DE LA MIXTECA

"CLASIFICACIÓN DE MASAS EN IMÁGENES DE MAMOGRAFÍAS UTILIZANDO REDES BAYESIANAS"

T E S I S

PARA OBTENER EL GRADO DE

MAESTRO EN TECNOLOGÍAS DE CÓMPUTO APLICADO

PRESENTA:

Rolando Pedro Gabriel

DIRECTORES DE TESIS:

Dr. Raúl Cruz Barbosa,

M.C. Verónica Rodríguez López.

HUAJUAPAN DE LEÓN, OAX., FEBRERO DE 2014

A mi familia.

Agradecimientos

Mis profundos agradecimientos van dirigidos a Dios por brindarme la fe y la consistencia para avanzar un escalón más en el ámbito profesional. Porque en el momento justo de visualizar imposible la culminación de este proyecto, puso en mi las fuerzas y la motivación para lograrlo. Porque antes de emprender esta etapa, me encontraba en un entorno reducido para un crecimiento profesional. Pero en Dios he encontrado la luz que me guía.

Agradezco especialmente a mi familia por el apoyo y la motivación que han inspirado en mi para lograr cruzar un sendero más en el amplio camino de la educación.

Con estas frases expreso mis más profundos y sinceros agradecimientos al Dr. Raúl Cruz Barbosa y a la M. C. Verónica Rodríguez López por la orientación, el seguimiento y la supervisión continua de la misma, pero sobre todo por la motivación y el apoyo que me han brindado.

A mis sinodales Dra. Lluvia Carolina Morales Reynaga, al Dr. Felipe de Jesús Trujillo Romero y al Dr. José Aníbal Arias Aguilar por su disposición en la revisión de este proyecto de tesis, así como los comentarios y observaciones para mejorar el documento.

A la Universidad Tecnológica de la Mixteca el haberme dado la oportunidad de recibir mi formación académica dentro de sus aulas y a mis profesores por el apoyo y conocimiento que de ellos recibí.

Agradezco al Consejo Nacional de Ciencia y Tecnología (CONACYT) por el apoyo económico otorgado para poder concluir mis estudios de maestría.

Contenido

Índice de figuras

Índice de cuadros

Capítulo 1

Introducción

El cáncer de mama es un problema de salud mundial que afecta principalmente a las mujeres y es la causa principal de muerte de mujeres jóvenes en los países desarrollados. Los países con mayor frecuencia de cáncer de mama son los que se ubican en la parte norte de América, una gran parte de los países europeos, Australia y Nueva Zelanda [7, 54, 71]. Mientras que en México, el problema del cáncer de mama se ha convertido en una situación difícil para la población de mujeres adultas, por razones económicas, sociales, tecnológicas y médicas [63]. Los datos históricos mencionan que el cáncer mamario ha registrado un aumento sustancial de 1950 a 2008, pasando de una tasa de 2 por cada 100 mil mujeres a 9 por cada 100 mil mujeres mexicanas. A partir del 2006, el cáncer de mama ocupa el primer lugar de mortalidad por neoplasia maligna entre las mujeres mexicanas de 25 años en adelante y el segundo lugar como causa de muerte entre las mujeres de 30 a 54 años [38]. Según el Instituto Nacional de Estadística y Geografía (INEGI), entre 2005 y 2009 la tasa estandarizada del cáncer de mama pasó de 17.9 a 10.8 fallecimientos por cada 100 mil mujeres de 25 años en adelante, mientras que para el 2010 se tiene un reporte de una tasa de mortalidad de 8.8 por cada 100 mil mujeres [17].

Uno de los métodos más viables, de alta precisión y de bajo costo, para la detección oportuna del cáncer de mama es la mamografía. La mamografía es un tipo particular de imágenes que se obtienen de la mama usando una baja dosis de rayos X. Estas imágenes permiten examinar la mama para detectar posibles microcalcificaciones, masas, distorsión en la arquitectura y asimetría bilateral [20]. Las microcalcificaciones y las masas son los indicadores más importantes de la presencia del cáncer [11]. Las microcalcificaciones

son pequeñas partículas blancas que no pertenecen a la anatomía regular de la mama y que se presentan en los ductos lactíferos, las cuales se pueden visualizar en las mamografías como pequeños cúmulos. Por otra parte, las masas se definen como una lesión que ocupa un espacio observable en al menos dos proyecciones diferentes. Las masas presentan mayor dificultad que las microcalcificaciones para ser descubiertas en las mamografías, debido principalmente a sus características y al tipo de tejido mamario [61].

CADx (*Computer-Aided Diagnosis*) es un tipo de sistema computacional que ha sido desarrollado para ayudar a los radiólogos en el análisis de mamografías. Estos sistemas tienen por objetivo apoyar a los radiólogos en la detección y diagnóstico de lesiones [61]. Un sistema CADx se compone de diversas etapas de procesamiento. Inicialmente, a la imagen de mamografía se le aplican métodos de preprocesamiento con la finalidad de mejorar su calidad y obtener mejores resultados en la segmentación. Después, se realiza la segmentación, que se refiere a la detección de regiones sospechosas en la mamografía. Esta etapa es muy importante para determinar la sensibilidad de todo el sistema, ya que el éxito de los algoritmos de clasificación depende de esta etapa. La siguiente etapa es la extracción y selección de características que permitirán diferenciar lesiones malignas y benignas. Por último, se realiza la clasificación de la lesión, como maligna o benigna [11].

El desarrollo de este proyecto se enfoca en la implementación de un clasificador de masas de mamografía utilizando una red Bayesiana. Se realizará un análisis sobre diversos tipos de redes Bayesianas y se comparará su desempeño con otro tipo de clasificador. Así mismo, se implementará la red Bayesiana que proporcione mejores resultados y un módulo para realizar la extracción de características de forma automática.

1.1. Planteamiento del Problema

El cáncer de mama es una de las enfermedades que se presenta con mayor frecuencia en la población femenina y que ocasiona terribles daños físicos, emocionales, sociales y en el caso más grave la muerte [38]. Para reducir esta amenaza, se requiere de la detección y el diagnóstico oportuno de este tipo de cáncer, sin embargo, estas son tareas difíciles a las que se enfrentan los radiólogos, ya que requieren de conocimiento y experiencia.

Una de las anomalías más difíciles de detectar y diagnosticar son las

masas [11]. Para diagnosticar masas, los radiólogos realizan una evaluación de sus características: forma, margen, tamaño y densidad. La forma de las masas son muy variadas y sus márgenes son difíciles de visualizar ya que su intensidad es similar al tejido normal. Otros factores que afectan el análisis de masas son su tamaño y su alta densidad. Dada la complejidad y variabilidad de las características de las masas, el diagnóstico de estas lesiones depende del nivel de experiencia del radiólogo [6].

La automatización del diagnóstico de masas a través de una computadora, debido a su complejidad, es un problema que continúa siendo investigado. Algunas de las técnicas que se han utilizado para resolver este problema son Análisis Discriminante Lineal (LDA), redes Neuronales, redes Bayesianas, Árboles de Decisión Binarios y Máquinas de Soporte Vectorial [6, 11]. Las redes Bayesianas destacan en aplicaciones médicas debido al hecho de que resaltan a través de su estructura relaciones causales entre síntomas y una enfermedad.

El objetivo de este proyecto es implementar una red Bayesiana que permita clasificar las masas en imágenes de mamografías como malignas o benignas. Es importante destacar, que este clasificador se evaluará con características de las masas que se obtendrán automáticamente. Por lo tanto, también se contempla implementar un módulo de extracción de características de las masas contenidas en imágenes segmentadas previamente (manual o automáticamente). Se investigarán diversos grupos de características que describan la forma, los márgenes, el tamaño y la densidad de las masas. El desempeño de la red Bayesiana se comparará con el de otro tipo de clasificador.

1.2. Justificación

Una de las razones que motiva el desarrollo de este proyecto es el rezago tecnológico para combatir problemas de salud en México, en este caso el cáncer de mama. A pesar de que este tipo de cáncer ocurre con mayor frecuencia en los países de alto nivel socioeconómico, los países con marginación y pobreza tienen mayor probabilidad de casos de defunciones [63]. Además, debido a factores sociales y a la falta de información sobre el cáncer de mama, la gran mayoría de casos son detectados a través de una exploración física y solamente algunos mediante análisis mamográficos [46].

Un análisis adecuado de mamografías permite detectar oportunamente el cáncer de mama, sin embargo, como ya se mencionó anteriormente, esta es una tarea difícil que sólo puede ser realizada por radiólogos especializados. Desafortunadamente, nuestro país no cuenta con el número necesario de radiólogos especializados para realizar interpretaciones mamográficas. Además, diversos estudios reportan que el desempeño de los radiólogos en las interpretaciones mamográficas es aproximadamente del 75 % [11, 61]. Este rendimiento se podría mejorar con la contribución de un sistema CADx, para los cuales se ha reportado un rendimiento entre el 80 y 90 % [47].

En este proyecto se plantea desarrollar un módulo de clasificación de masas que se pretende, forme parte de un sistema CADx, que apoye a los radiólogos en el diagnóstico de masas en imágenes mamográficas.

1.3. Hipótesis

Es posible la clasificación de masas utilizando redes Bayesianas y alcanzar un desempeño similar o mayor al de un radiólogo.

Otra hipótesis del proyecto de tesis se basa en que "la diferencia entre patrones característicos de grupos" es útil para la clasificación de masas, esto significa que se extraerán parámetros característicos de las mamografías para compararlos con los grupos de características de lesiones típicamente malignas y benignas.

1.4. Trabajos relacionados

Actualmente las redes Bayesianas son ampliamente utilizadas en diversos campos de la investigación debido a que proporcionan métodos sistemáticos para estructurar información probabilística, además son una base fundamental para la toma de decisiones [16]. A continuación se mencionan algunos trabajos relacionados con este proyecto de tesis.

Diversas son las redes Bayesianas que se han propuesto para el diagnóstico del cáncer de mama. En [34], [59] y [74] se proponen redes Bayesianas con estructuras definidas a partir del conocimiento de expertos, y entrenadas con datos proporcionados por radiólogos especializados. Las características corresponden al historial clínico de la paciente (edad, número de familiares, la edad del primer hijo, edad de la menarquía y la biopsia previa), hallazgos

físicos (dolor y secreción del pezón) y hallazgos mamográficos (propiedades de las masas y calcificaciones). Otra red Bayesiana entrenada a partir de descriptores proporcionados por radiólogos especializados se propone en [25]. Las características son descriptores BI-RADS (Sistema de informes y registros de datos de imagen de la Mama o *Breast Imaging Reporting and Data System*) de masas y calcificaciones; y para obtener la topología de la red, aplicaron métodos de aprendizaje automático.

Uno de los trabajos que más se relaciona con el proyecto que se pretende realizar, es el de [62]. En esta investigación, se realizó un análisis de diversos tipos de redes Bayesianas y una Máquina de Soporte Vectorial (Support Vector Machine - SVM). Los modelos propuestos permiten clasificar el tumor en tres clases, benigno, cáncer in-situ y cáncer invasivo. Para construir los modelos utilizaron 81 características obtenidas de forma automática de las regiones sospechosas de contener una lesión. Entre los tipos de redes Bayesianas que se analizan están Naïve Bayes, Naïve Bayes aumentado a árbol (TAN) y Naïve Bayes aumentado a bosque (FAN).

1.5. Objetivos

1.5.1. Objetivo General

Desarrollar una red Bayesiana para la clasificación de masas malignas y benignas en imágenes de mamografía.

1.5.2. Objetivos Específicos

- Implementar un módulo en Java para la obtención de características de las lesiones consideradas como masas.
- Seleccionar un conjunto de características que describan las propiedades de la masa.
- Construir y analizar diversos tipos de redes Bayesianas e implementar en Java la mejor.
- Comparar el desempeño de la red Bayesiana con otro tipo de clasificador.

1.6. Metas

- Investigar sobre el problema del cáncer de mama y las herramientas para su detección y diagnóstico (CAD y CADx).
- Analizar e implementar en el lenguaje de programación Java diversos métodos para describir la región de una imagen.
- Aplicar diversos métodos para la selección de un conjunto de características óptimas que se utilizarán en la construcción de las redes Bayesianas.
- Investigar diversos tipos de redes Bayesianas como Naïve Bayes, Naïve Bayes aumentado a árbol (TAN), Clasificadores Bayesianos K-Dependientes (KDB) y Naïve Bayes aumentado a bosque (FAN).
- Construir y comparar diversas redes Bayesianas en el entorno de Matlab.
- Implementar en el lenguaje de programación Java la red Bayesiana con mejor rendimiento.
- Comparar el desempeño de la red Bayesiana implementada con otro clasificador.
- Publicación de, al menos, un artículo arbitrado.

1.7. Metodología

Para abordar el tema del cáncer de mama se realizará un estudio detallado sobre el dominio de este tipo de cáncer y los métodos que se han utilizado para abordar este tópico en el ámbito computacional. Como ya se mencionó en la introducción, existen tres indicadores importantes (masas, microcalcificaciones y distorsión de la arquitectura) para determinar la presencia del cáncer de mama en imágenes de mamografía. Este proyecto de investigación está enfocado al análisis de masas.

Las imágenes de mamografía que se utilizarán en este proyecto de tesis se obtendrán de la base de datos mini-MIAS [67]. Mini-MIAS es una base de datos pública que ha sido utilizada en diversas investigaciones. Las imágenes que se proporcionan están en formato PGM con una resolución de 8 bits por cada píxel. Esta base de datos contiene información sobre la ubicación, el tipo (circunscritas, mal definidas y espiculadas) y la clasificación de la lesión

(benigna o maligna) [67].

A fin de construir el módulo de extracción de características, se realizará un análisis sobre diversos métodos de descripción de regiones de imágenes digitales, que permitan obtener las características de forma, textura e intensidad de las masas. Debido a que se planea que el sistema CADx a desarrollarse, en un futuro, se pueda ejecutar desde un sitio web, este módulo se implementará en el lenguaje de programación Java.

El proceso de construcción del clasificador conlleva una investigación exhaustiva de diferentes tipos y topologías de redes Bayesianas. Se realizará el aprendizaje de múltiples redes Bayesianas utilizando el toolbox para Matlab de Kevin Murphy (*Bayesian Network Toolbox BNT*) [50]. La mejor red Bayesiana formará parte del sistema de diagnóstico de masas, el cual se implementará en Java. Finalmente, la red Bayesiana seleccionada se evaluará con, al menos, un método de clasificación diferente.

Capítulo 2

Diagnóstico de cáncer de mama mediante mamografías

En este capítulo se proporciona una breve explicación de la anatomía de la mama y los tipos de tumores que en esta zona se pueden desarrollar. Así mismo, se presenta una descripción de las imágenes de mamografía y los hallazgos que se pueden detectar en éstas. Finalmente, se reseñan los sistemas de detección asistido por computadora (CADe) y los sistemas de diagnóstico asistido por computadora (CADx).

2.1. Cáncer de mama

El cáncer de mama es la formación de un tumor maligno que se desarrolla a partir de la proliferación acelerada y desordenada de células en la mama. Es el problema de salud pública más significante en el mundo y es el tipo de cáncer más frecuentes entre las mujeres, con mayor tasa de mortalidad en mujeres de 25 años en adelante. El riesgo de desarrollar el cáncer de mama en las mujeres en países desarrollados se ha calculado alrededor de 1 de cada 8 [8]. A pesar de que su presencia sigue aumentando entre la población femenina, la mortalidad ha ido disminuyendo debido en gran medida a la detección oportuna y a las mejoras en el tratamiento [15]. Sin embargo, la alta incidencia, la complejidad y el costo económico del tratamiento para esta enfermedad hacen que el cáncer de mama sea uno de los problemas de salud más relevantes en nuestra sociedad [72].

2.1.1. Anatomía de la mama

La mama se encuentra ubicada sobre el músculo pectoral que lo separa del resto del cuerpo. La mama se compone de tres tipos de tejidos: fibroso, graso y del parénquima. El tejido del parénquima se compone de 15 a 20 lóbulos, los cuales están separados por paredes fibrosas y divididos en secciones más pequeñas llamadas lobulillos. Los lóbulos están conectados al pezón de la mama a través de una red de conductos. Los conductos comienzan a partir de los lóbulos de forma muy delgada, hasta unirse con el pezón, en esta parte los conductos son más grandes. Por lo tanto, agrupando el tejido glandular (lóbulos y conductos) junto con el tejido fibroso se tiene el tejido fibroglandular [56].

2.1.2. Tumores benignos

Muchos de los tumores que se desarrollan en la glándula mamaria son benignos, debido a formaciones fibroquísticas. Un quiste es una agrupación de líquido y la fibrosis es el desarrollo anormal del tejido conjuntivo, con frecuencia forman una masa. Cuando los quistes son grandes resultan ser dolorosos pero no son peligrosos y no se propagan fuera de la mama hacia otros órganos. Los tumores benignos están relacionados, generalmente con factores genéticos. Si el tumor benigno es grande, puede cambiar el tamaño y la forma de la mama. Si crece hacia el tejido de los conductos mamarios, puede causar secreción anormal del pezón. Los médicos pueden recomendar una extirpación mediante cirugía, dependiendo del tipo, tamaño y de la cantidad de tumores benignos [13].

2.1.3. Tumores malignos

Los tumores malignos indican la presencia del cáncer. Estos tumores se componen de las células que invaden y dañan los tejidos y órganos cercanos, así mismo, estas células malignas se pueden desprender del tumor maligno y entrar en la circulación o sistema linfático. Existen cuatro diferencias importantes en el crecimiento de las células cancerosas a las células normales, las cuales son [13]:

- Autonomía: Crecimiento desenfrenado de las células malignas.
- Clonación: Una única célula progenitora que prolifera y origina un clon de células malignas.
- La anaplasia: Ausencia de diferenciación normal y coordinada.

- La metástasis: Capacidad de crecer y diseminarse a otras partes del cuerpo.

Cuando el tumor maligno comienza a desarrollarse, se le denomina cáncer de mama *in situ* y cuando el tumor ya ha avanzado, de tal manera que las células malignas se han propagado considerablemente, se le denomina cáncer de mama invasivo.

Cáncer in situ

Se le denomina cáncer de mama *in situ*, debido a que las células malignas se encuentran en una etapa inicial de desarrollo y no se han propagado al tejido adiposo de la mama ni a otros órganos del cuerpo. Es muy probable la cura para este tipo de cáncer, si el tumor maligno se extirpa en su totalidad. Existen dos tipos de carcinoma *in situ* [13]:

- *Carcinoma lobular in situ:* Se origina en los lobulillos, pero no ha crecido a través de las paredes del lobulillo.

- *Carcinoma ductal in situ:* Las células cancerosas dentro de los conductos no se propagan a través de las paredes de los conductos hacia el tejido adiposo de la mama.

Cáncer invasivo

En este caso, las células anormales se han propagado más allá del lugar en el que comenzaron. El cáncer de mama invasivo puede originarse en los conductos o en los lobulillos [13].

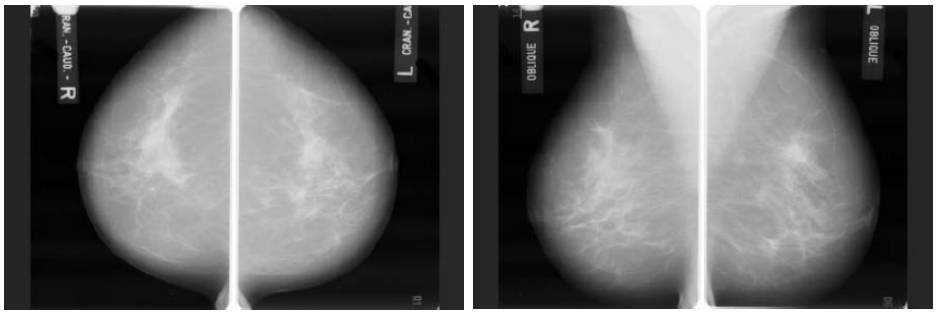
- *Carcinoma ductal invasivo:* Se origina en un conducto, penetra en sus paredes y se propaga al tejido de la mama, las células malignas pueden propagarse a los canales linfáticos o a los vasos sanguíneos accediendo a otros órganos del cuerpo, es el más frecuente de los cánceres, aproximadamente el 75 %.

- *Carcinoma lobular invasivo:* Se origina en los lobulillos y se propaga por sus paredes al tejido adiposo luego a los canales linfáticos y al torrente sanguíneo.

2.2. Mamografía

La mamografía (también llamada mastografía) es una imagen plana obtenida de la mama a través de una baja dosis de rayos X y es considerado el estudio por excelencia para la detección oportuna del cáncer de mama [40]. En las mamografías el tejido fibroglandular es altamente absorbente de los rayos X, mientras que el tejido graso no lo es, por lo que la región del tejido fibroglandular de la mama aparece más brillante en las mamografías. Las imágenes de mamografías se pueden obtener de diferentes ángulos de la mama, las dos proyecciones más comunes son Cráneo-Caudal (CC) y Media-Lateral-Oblicua (MLO) (ver Figura 2.1). La CC es una proyección vertical, vista superior a inferior de la mama, en la que se puede observar con mayor precisión el tejido medial. La MLO es una proyección a 45 ° con respecto del eje de simetría del cuerpo de la paciente, la ventaja de esta proyección es que casi toda la mama es visible, a menudo incluye los ganglios linfáticos [62, 18]. En las imágenes de mamografías se pueden visualizar hallazgos que se encuentran entre los componentes del tejido de la mama, tales como, calcificaciones, masas y distorsión de la arquitectura. Este tipo de estudio proporciona alta sensibilidad y especificidad, incluso, tumores pequeños y microcalcificaciones pueden ser detectados [45].

El análisis de mamografías producidas por una película de rayos-X, tienen mayor sensibilidad y especificidad para la detección del cáncer de mama que otras técnicas de diagnóstico no invasiva actualmente en uso. En los últimos años, la producción de mamografías han evolucionado a las aplicaciones de mamografías digitales, con el objetivo de mejorar algunos de los problemas de rendimiento y de calidad relacionadas con las imágenes generadas por películas de rayos-X. En la mamografía digital la película de rayos-X es reemplazada por detectores en estado sólido que transforman los rayos X en señales eléctricas, produciendo imágenes de las mamas que pueden verse a través de una computadora o ser impresas en una película especial similar a las mamografías convencionales. Algunas de las ventajas que presentan las mamografías digitales ante las mamografías convencionales, es que se pueden guardar y recuperar electrónicamente, lo que facilita las consultas a distancia con otros especialistas en mamografías. Dado que las imágenes pueden ser ajustadas por el radiólogo, las diferencias sutiles entre los tejidos pueden visualizarse con mayor facilidad. Además, puede ayudar a reducir el número de procedimientos de seguimiento y presenta mayores beneficios para las mujeres con tejido denso (aunque es menos eficaz en mujeres con mama adiposo) [64].



(a) (b)

Figura 2.1: Las dos proyecciones más comunes de la mama: a) Cráneo-Caudal (CC) y b) Media-Lateral-Oblicua (MLO).

2.3. Sistema BI-RADS

El Colegio Americano de Radiología (*American College of Radiology ACR*) desarrolló un estándar para clasificar los hallazgos mamográficos denominado BI-RADS (*Breast Imaging Reporting and Data System*) con la finalidad de ofrecer a los radiólogos una forma de categorizar los resultados y crear un plan de seguimiento [57, 23]. A continuación se describen las categorías y las recomendaciones que contempla este estándar:

Bi-Rads 0: Evaluación adicional. Se considera una categoría incompleta, se recomienda realizar estudios adicionales (magnificación, ecografía, etc.) o la comparación con mamografías anteriores.

Bi-Rads 1: Mama normal (ningún hallazgo). Se recomienda seguimiento a intervalo normal.

Bi-Rads 2: Benigna. Normal, pero existen hallazgos benignos, se recomienda seguimiento a intervalo normal.

Bi-Rads 3: Probablemente benigna. Hallazgos con una probabilidad $<2\,\%$. Se describen tres hallazgos específicos: nódulo sólido circunscrito no calcificado, asimetría focal y microcalcificaciones puntiformes agrupadas. En esta categoría no se contemplan lesiones palpables, sin embargo, se recomienda realizar una evaluación completa a través de proyecciones adicionales, como la ecografía o la comparación con estudios previos.

Bi-Rads 4: Anormalidad sospechosa. Incluye aquellas lesiones que van a requerir intervencionismo, tiene un rango de malignidad del 2-95 %. Se sugiere una división en tres subcategorías [3]:

- **4a.** Muy baja probabilidad de malignidad, en esta categoría se pueden integrar una masa sólida palpable parcialmente circunscrita, un quiste palpable o un absceso mamario.
- **4b.** Probabilidad intermedia de malignidad, en esta categoría se puede incluir una masa parcialmente circunscrita con márgenes mal definidos.
- **4c.** Riesgo moderado de malignidad, en esta categoría se pueden clasificar las masas irregulares con márgenes mal definidos o las microcalcificaciones. El resultado que se espera en esta categoría es maligno.

Bi-Rads 5: Altamente sugestiva de malignidad. Hallazgos típicamente malignos, con una probabilidad >95 %. Es indispensable un estudio de histopatología.

Bi-Rads 6: Malignidad conocida. Lesión con malignidad ya demostrada mediante biopsia.

2.3.1. Calcificaciones

Las calcificaciones son pequeños depósitos de minerales (calcio) que se encuentran en la mama en regiones de alta intensidad, y se pueden visualizar como manchas en la mamografía. Existen dos tipos de calcificaciones: microcalcificaciones y macrocalcificaciones. Las macrocalcificaciones son depósitos de calcio de forma dispersa y gruesa. Las microcalcificaciones pueden encontrarse de forma aislada o incrustadas en una masa. El tamaño de las microcalcificaciones van desde 0.1 hasta 1.0 mm con un diámetro medio de 0.5 mm. Del 30-50 % de los cánceres no palpables se detectan inicialmente debido a la presencia de grupos de microcalcificaciones [64].

2.3.2. Distorsión de la arquitectura

Se refiere a un cambio anormal de la mama formándose lesiones finas y espiculadas que no están asociadas a la presencia de una masa. La distorsión de la arquitectura hace referencia a la distorsión del parénquima de la mama pero sin la presencia de masas ni aumento en la densidad. Es el tercer hallazgo más común en la mamografía que está asociado a estados de cáncer aún no palpables [29].

2.3.3. Masas

La masa se define como una lesión que ocupa un espacio y es visible en al menos dos proyecciones (CC y MLO). Si sólo se visualiza en una

proyección, entonces se considera como densidad, hasta que no se compruebe su tridimensionalidad. Con mayor frecuencia el cáncer de mama se presenta como una masa con o sin la presencia de calcificaciones. Una masa representa un quiste que se forma a partir de una colección de líquido no canceroso. La dificultad para decidir si es una masa, se debe a la similitud de sus intensidades con las del tejido normal. De acuerdo al sistema BI-RADS, las masas se caracterizan por su forma, contorno y densidad, como se muestra en el Cuadro 2.1 [64, 57].

Forma	**Contorno**	**Densidad**
Redondo	Circunscrito	Hiperdenso
Ovalado	Microlobulado	Isodenso
Lobulado	Ocultos	Hipodenso sin grasa
Irregular	Mal definido	Hipodenso con grasa
	Espiculado	

Cuadro 2.1: Posibles formas, contornos y densidades de una masa considerados por el sistema BI-RADS.

Forma

Las posibles formas de una masa se muestran en la Figura 2.2 y son [3]:

- Redonda: Una masa que es esférica, circular o globular.
- Ovalada: Una masa que es elíptica o en forma de huevo.
- Lobulada: Una masa que tiene contornos ondulados.
- Irregular: La forma de la lesión no puede caracterizarse por ninguna de las terminologías descrita anteriormente.

Contorno

El contorno o márgenes modifican los límites de las masas, como se observa en la Figura 2.3, estos pueden ser [3, 64]:

- Circunscritos: Contornos bien definidos, claramente demarcados, con transición abrupta entre la lesión y el tejido adyacente.

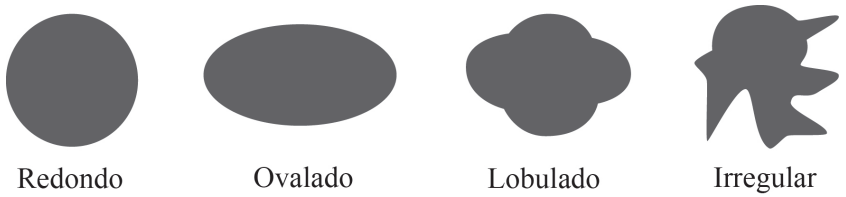


Figura 2.2: Posibles formas de una masa.

- Microlobulados: Este tipo de márgenes se caracterizan por presentar pequeñas ondulaciones.
- Oscurecido: Se asigna esta categoría cuando los bordes de la masa están ocultos parcialmente por superposición o por tejido adyacente normal, que impide definirlos.
- Mal definidos: La mala definición de los bordes se debe a infiltración por la lesión y no puede atribuirse a tejido normal superpuesto.
- Espiculados: La lesión se caracteriza por líneas que se irradian a partir de los márgenes.

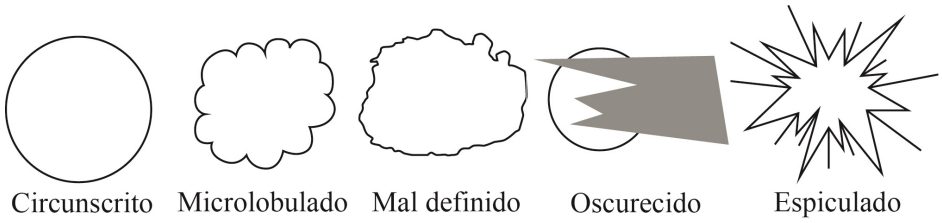


Figura 2.3: Posibles márgenes de una masa.

Densidad

La densidad de una masa se refiere a cómo se comporta el tejido de la masa con respecto al tejido presente en las estructuras densas de la mama (parénquima no graso, conductos y vasos sanguíneos) [65]. La densidad se divide en alta densidad, igual densidad, baja densidad (menor atenuación, pero no contiene grasa) y de contenido graso [3].

2.4. Detección y diagnóstico asistido por computadora

El análisis de imágenes mamográficas es una tarea muy importante para conocer si existen cambios anormales en la glándula mamaria. Sin embargo, para los radiólogos analizar estos tipos de imágenes es una tarea compleja. Para apoyar a los especialistas en el análisis de estos tipos de imágenes, se han desarrollado dos tipos de sistemas computacionales, el sistema de detección asistido por computadora (*Computer Aided Detection CAD o CADe*) y el sistema de diagnóstico asistido por computadora (*Computer Aided Diagnosis CADx*).

2.4.1. CADe

Los sistemas CADe o CAD se utilizan para ayudar a los radiólogos en la detección de anormalidades en mamografías. Son utilizadas como una segunda opinión en el análisis de mamografías, después de que el radiólogo realiza una primera interpretación [64, 5]. La tarea de estos sistemas es únicamente indicar las áreas a donde posiblemente puede existir una lesión, usando marcas especiales. El trabajo de delimitar el área de la lesión para que pueda ser caracterizada y diagnosticada, queda bajo la responsabilidad del radiólogo.

Las tareas que realiza un sistema CADe sobre imágenes de mamografías son:

- Preprocesamiento: Etapa en la cual se debe reducir el ruido y mejorar la calidad de la imagen. En algunos sistemas, en esta etapa es a donde se elimina el fondo de la imagen y el músculo pectoral.
- Segmentación: Etapa en la cual se localizan y se busca aislar regiones sospechosas de pertenecer a una anormalidad.
- Extracción de características: Etapa en la que se obtienen las características de las regiones sospechosas obtenidas en la etapa anterior.
- Clasificación: Etapa en la que se clasifican las regiones sospechosas utilizando sus características. Esta etapa la aplican algunos sistemas CADe a fin de validar que la región detectada corresponde a una lesión y no al tejido mamario normal.

Existen diversos sistemas comerciales tipo CAD que han sido aprobados por la FDA (Food and Drug Administration, USA) para auxiliar en la detección de anormalidades en mamografías. Algunos de estos sistemas son Second Reader, R2 Technology's Image Checker, MammoReader y Second Look [61]. El rendimiento de un sistema de detección comercial es del 95 % en sensibilidad [61].

2.4.2. CADx

Los sistemas CADx son herramientas empleadas para apoyar en el diagnóstico de lesiones. Las tareas de los sistemas CADx son mucho más complejas que la de los CAD's. Estos sistemas, además de localizar lesiones, deben de caracterizarlas y determinar su grado de malignidad o benignidad. Debido a esta complejidad en sus funciones, aún no existen sistemas CADx comerciales. De acuerdo a algunas investigaciones [6, 61] estos sistemas tienen un desempeño del 86 % en sensibilidad.

Las etapas de un sistema CADx son las siguientes (Figura 2.4) [11].

- Preprocesamiento de las imágenes: Tiene la finalidad de resaltar las diferencias entre los cambios anormales y el tejido normal de la mama.
- Segmentación: Trata de aislar del resto de la imagen las regiones sospechosas que podrían contener alguna lesión. Este proceso es muy importante, porque determina la sensibilidad del sistema completo.
- Extracción de características: Consiste en obtener descriptores que describan las propiedades de la región de interés.
- Selección de características: El espacio de características puede ser muy grande y complejo, debido a la gran variabilidad que puede darse tanto en el tejido sano como en el canceroso. Por lo tanto, se recomienda seleccionar un subconjunto de características que permitan un buen funcionamiento del clasificador.
- Clasificación: Una vez seleccionadas las características, estas deben ser clasificadas para determinar el grado de sospecha de malignidad.

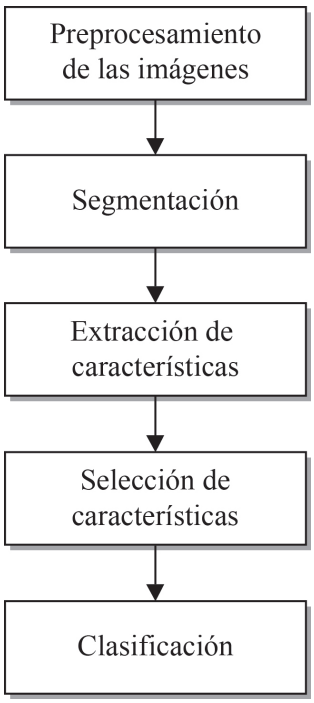


Figura 2.4: Etapas principales de un sistema de diagnóstico asistido por computadora CADx.

Capítulo 3

Extracción de características

Una región de una imagen puede revelar diferentes tipos de características de interés, esto se debe a que cada píxel puede almacenar diferentes tonalidades de colores que describen un objeto. El área de procesamiento de imágenes proporciona diversos métodos que permiten estudiar las propiedades que describen una región de la imagen. En el presente capítulo se describen algunos métodos para obtener las características de intensidad, forma, textura y ubicación de las masas presentes en una imagen de mamografía.

3.1. Descriptores de intensidad

Estos descriptores solo consideran la intensidad, $p(i, j)$, de cada píxel en la región, por lo que son los más simples y rápidos de obtener. Algunos descriptores de este tipo son [11, 30]:

Media. Mide el promedio de intensidades de los píxeles de la región.

$$\mu = \frac{1}{NM} \sum_{i=1}^{N} \sum_{j=1}^{M} p(i, j) \tag{3.1}$$

Mediana. Ordena las intensidades de los píxeles, proporcionando el píxel medio cuando el número de píxeles es impar, de lo contrario proporciona el valor de los dos píxeles centrales divididos entre 2.

Varianza. Mide la disimilitud que hay entre la distribución de las intensidades de los píxeles con respecto a la media.

$$s = \frac{1}{NM} \sum_{i=1}^{N} \sum_{j=1}^{M} (p(i,j) - \mu)^2 \tag{3.2}$$

Asimetría. Mide la distribución de las intensidades de los píxeles. Se puede visualizar una distribución simétrica, cuando se visualiza la misma cantidad de intensidad de píxeles a la izquierda y a la derecha, con respecto al punto central. Sin embargo, cuando la distribución se concentra hacia la izquierda se considera como negativa, y positiva cuando la distribución se concentra hacia la derecha.

$$asimetria = \frac{1}{NM\sigma^2} \sum_{i=1}^{N} \sum_{j=1}^{M} (p(i,j) - \mu)^3 \tag{3.3}$$

Curtosis. Estima la forma de la distribución de las intensidades de los píxeles en la región.

$$curtosis = \frac{1}{NM\sigma^4} \sum_{i=1}^{N} \sum_{j=1}^{M} (p(i,j) - \mu)^4 - 3 \tag{3.4}$$

Contraste. Se define como la diferencia relativa en la intensidad promedio de los niveles de gris de los píxeles dentro de la región de interés $E(I)$, y la intensidad promedio de los niveles de gris de los píxeles que se encuentran dentro de la banda que cubre la región de interés $E(O)$, como se muestra en la Figura 3.1b [69].

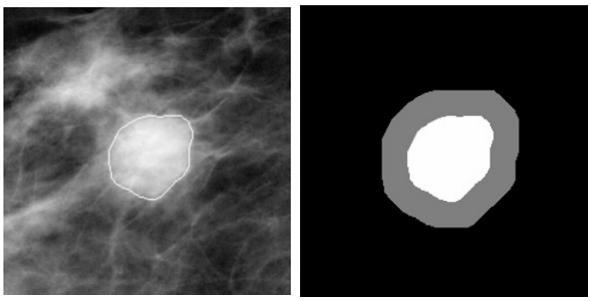
$$contraste = E(I) - E(O) \tag{3.5}$$

Para construir la banda que se muestra en la Figura 3.1b, se considera el radio efectivo de la región de interés mediante la siguiente ecuación:

$$R = \sqrt{\frac{Area}{\pi}} \tag{3.6}$$

Y el ancho de la banda se obtiene mediante la siguiente ecuación:

$$anchoBanda = 0.6 * R \tag{3.7}$$



(a) (b)

Figura 3.1: Elementos para el cálculo del contraste de la masa. a) Imagen original con la masa o región de interés (la región delimitada). b) Región de interés en blanco, y en gris, la banda que cubre esta región.

3.2. Descriptores de forma

Las características de forma o morfológicas, son consideradas los descriptores más importantes de un objeto [4]. Para describir una región inscrita en una imagen, se deben contemplar dos propiedades importantes, la primera es mediante sus características externas (su contorno) y la segunda a través de sus características internas (los píxeles que comprenden la región) [30].

3.2.1. Descriptores simples

Algunos descriptores simples de forma son [30]:

Área (A). Se define como el número de píxeles contenidos dentro de los límites de la región.

Perímetro (P). Es el número de píxeles que se encuentran en el contorno de la región.

Compacidad. Es un descriptor de forma que indica qué tan compacto es un objeto. La compacidad representa la rugosidad del límite en relación al área del objeto, y se obtiene mediante la ecuación 3.8:

$$compacidad = \frac{P^2}{4\pi A} \tag{3.8}$$

3.2.2. Longitud Radial Normalizada (LRN)

La Longitud Radial Normalizada (LRN) se define como la distancia Euclidiana normalizada, $r(i, j)$, entre el centroide de la región hacia uno de los

puntos de su contorno, como se muestra en la Figura 3.2. La distancia se normaliza al considerar la máxima distancia Euclidiana que existe hacia un punto del contorno [68, 1]. Las mediciones que se obtienen a partir de este método son las siguientes [1, 37, 19, 12]:

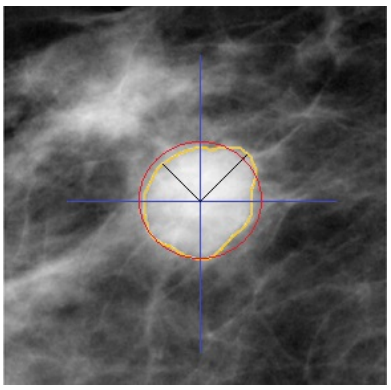


Figura 3.2: Distancia radial normalizada.

Media. Es una medida que describe el promedio de la distancia que existe a partir del centroide hacia un punto del contorno.

$$r_{avg} = \frac{1}{P}\sum_{i=1}^{p} r(i) \tag{3.9}$$

Desviación Estándar. Es una buena medición para la irregularidad del contorno de la región. Este valor aumenta cuando existe una mayor espiculación e irregularidad.

$$\sigma = \sqrt{\frac{1}{P}\sum_{i=1}^{p}(r(i) - r_{avg})^2} \tag{3.10}$$

Entropía. Es una medida probabilística que se obtiene a partir del histograma de la LRN, como se muestra a continuación.

$$E = -\sum_{i=1}^{N_{bins}} P_k log P_k \tag{3.11}$$

Donde P_k es la probabilidad de que la LRN se encuentre entre $r(i)$ y $r(i+1)+1/N_{bins}$, donde N_{bins} es el número de intervalos del histograma de la LRN, que varía en el intervalo $[0,1]$.

Índice de área. Este indicador mide el porcentaje del área de la región de interés que se encuentra fuera de una región circular.

$$IA = \frac{1}{r_{avg} * P} \sum_{i=1}^{p} (r(i) - r_{avg}) \tag{3.12}$$

Rugosidad del contorno. Se utiliza para describir el grado de espiculación de la región de interés.

$$RC = \frac{1}{P} \sum_{i=1}^{p} |r(i) - r(i+1)| \tag{3.13}$$

Cruce por cero. Es el conteo del número de veces de que $r(i) > r_{avg}$.

$$CC = \sum_{i=1}^{p} h(r(i), r_{avg}) \tag{3.14}$$

donde $h(r(i), r_{avg})$ se obtiene de la siguiente manera:

$$h(r(i), r_{avg}) = \begin{cases} 1 & \text{para } r(i) > r_{avg} \\ 0 & \text{de otra manera} \end{cases} \tag{3.15}$$

3.2.3. Momentos invariantes de Hu

Los momentos de una imagen digital describen la geometría de una región plana basándose en el tamaño, la posición, la orientación y la forma. Los momentos invariantes se obtienen independientemente de transformaciones geométricas que haya sufrido la imagen, como rotación, traslación y cambios de escala de los objetos. La teoría de los momentos invariantes fue introducida por Ming-Kuei Hu, por medio del teorema fundamental de

momentos invariantes [53].

Para el análisis de forma, se calculan los siete momentos de Hu a partir de los momentos geométricos, centrales y centrales normalizados. Los momentos geométricos se obtienen a partir de una función $f(x, y)$ como la intensidad del punto (x, y) en una región. El momento de orden $(p + q)$ para la región se define como:

$$m_{pq} = \sum_x \sum_y x^p y^q f(x, y) \tag{3.16}$$

Los momentos centrales son útiles para reconocer una imagen independiente de su ubicación en un eje de coordenadas. El momento de orden central de orden $(p + q)$ viene dado por:

$$\mu_{pq} = \sum_x \sum_y (x - \bar{x})^p (y - \bar{y})^q f(x, y) \tag{3.17}$$

donde

$$\bar{x} = \frac{m_{10}}{m_{00}} \qquad \bar{y} = \frac{m_{01}}{m_{00}} \tag{3.18}$$

Y la normalización de los momentos centrales de orden $(p+q)$, se denota por η_{pq} y se define como:

$$\eta_{pq} = \frac{\mu_{pq}}{\mu_{00}^{\gamma}} \qquad donde \qquad \gamma = \frac{p+q}{2} + 1 \qquad para \qquad (p+q) = 2, 3, ... \tag{3.19}$$

Los siete momentos invariantes de Hu se pueden obtener usando únicamente los momentos centrales normalizados de órdenes 2 y 3 [30]:

$$\phi_1 = \eta_{20} + \eta_{02} \tag{3.20}$$

$$\phi_2 = (\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2 \tag{3.21}$$

$$\phi_3 = (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2 \tag{3.22}$$

$$\phi_4 = (\eta_{30} + \eta_{12})^2 + (\eta_{21} - \eta_{03})^2 \tag{3.23}$$

$$\begin{aligned}\phi_5 = (\eta_{30} - \eta_{12}) + (\eta_{30} + \eta_{12})\big[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2\big] + \\ (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03})\big[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2\big]\end{aligned} \tag{3.24}$$

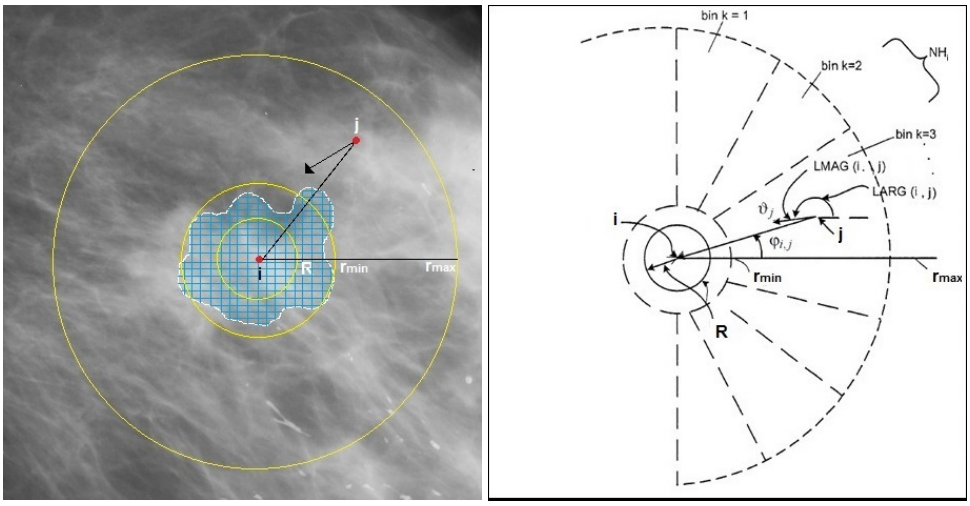
$$\begin{aligned}\phi_6 = (\eta_{20} - \eta_{02})\big[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2\big] \\ +4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03})\end{aligned} \tag{3.25}$$

$$\begin{aligned}\phi_7 = (3\eta_{21} - \eta_{03})\eta_{30} + \eta_{12})\big[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2\big] + \\ (3\eta_{21} - \eta_{30})(\eta_{21} + \eta_{03})\big[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2\big]\end{aligned} \tag{3.26}$$

3.2.4. Patrones estrellados

Un patrón estrellado es un conjunto de líneas rectas que apuntan hacia un píxel en particular. Este tipo de patrones permite describir la probabilidad de que una lesión mamográfica resulte maligna [62]. Para obtener patrones estrellados en una imagen, se utiliza el mapa de orientaciones de la imagen, el cual se puede obtener a partir del gradiente de la imagen [53].

Una de las características que puede ayudar a describir un patrón estrellado es $f1$. $f1$ es una medida normalizada de la cantidad de píxeles que tienen una orientación lineal hacia el píxel central i y que se ubican dentro de una región circular con radio R, donde i representa el centro de la región de interés. Para estimar esta característica, se analizan los píxeles que se ubican dentro de la vecindad determinada por r_{min} y r_{max}, como se muestra en la Figura 3.3a. Esta vecindad se divide en k fracciones y en cada fracción se determina el número de píxeles que tienen una orientación lineal



(a) (b)

Figura 3.3: Cálculo del descriptor f1. a) Píxeles en la vecindad circular que se analizan para realizar el cálculo. b)División de la vecindad circular y relación de los parámetros r_{min}, r_{max} y R.

hacia el centro de la región de interés i, como se ejemplifica en la Figura 3.3b [62, 69]. En este trabajo, r_{max} se considera como la distancia mínima que existe entre el píxel central i con los límites de la imagen original (izquierda, derecha, superior e inferior), $r_{min} = \frac{r_{max}}{4}$ y $R = \frac{r_{min}}{2}$.

Formalmente, $f1$ se calcula mediante la siguiente ecuación,

$$f_{1,i} = \frac{n_i - pN_i}{\sqrt{N_i p(1-p)}} \tag{3.27}$$

Donde N_i representa el número de píxeles j que se encuentran dentro de la vecindad determinada por r_{min} y r_{max}, es decir, píxeles j con una distancia al píxel central i, $r_{i,j} \in [r_{min}, r_{max}]$. Y n_i se obtiene a partir de la sumatoria $n_{i,k}$ por cada una de las k fracciones.

$$n_{i,k} = \sum_{j \in N_{i,k} \cap S} h(\vartheta_j, \varphi_{i,j}, r_{i,j}) \tag{3.28}$$

$h(\vartheta_j, \varphi_{i,j}, r_{i,j})$ es una función que determina si un píxel j está apuntando hacia un píxel central i,

$$h(v_j, \varphi_{i,j}, r_{i,j}) = \begin{cases} 1 & \text{para } |\varphi_{i,j} - \vartheta_j| < \frac{R}{r_{i,j}} \\ 0 & \text{de otra manera} \end{cases} \tag{3.29}$$

Por otra parte, p de la ecuación (3.27) determina la probabilidad media de que un píxel j que se ubica dentro de la vecindad determinada por r_{min} y r_{max} esté apuntando hacia el píxel central i.

$$p = \frac{2}{\pi N_i} \sum_{j \in N_{i,k} \cap S} \frac{R}{r_{i,j}} \tag{3.30}$$

Otro descriptor de un patrón estrellado es $f1$ promedio. Para calcular este descriptor se consideran a cada uno de los píxeles de la región de interés como el píxel central y se estima $f1$ para cada uno de estos píxeles. Finalmente, se realiza una sumatoria de las $f1$ calculadas, y se divide entre el tamaño de la región de interés.

Otra de las características que se obtiene para describir un patrón estrellado es $f2$. Este descriptor ayuda a determinar si en la región existen espículas con diversas orientaciones. Mientras mayor sea el número de espículas con diferentes orientaciones, mayor será la probabilidad de la presencia de un patrón estrellado en la región. Esta medición se obtiene al dividir la vecindad circular en K sectores circulares y estimando en qué proporción el descriptor $f1$ estrellado se distribuye uniformemente en cada uno de los sectores. Formalmente, $f2$ se obtiene de la siguiente manera:

$$f2 = \frac{n_+ - K'/2}{\sqrt{K'/4}} \tag{3.31}$$

Donde n_+ es el número de veces que $n_{i,k}$ es mayor a la mediana de los píxeles que apuntan hacia el centro de la región i para cada fracción k. La mediana se obtiene mediante la siguiente ecuación,

$$mediana = n_{i,k} * p \tag{3.32}$$

3.2.5. Textura lineal

En general, las densidades mamográficas malignas tienen un aspecto irregular, en muchos de los casos se encuentran rodeadas de una radiación de líneas espiculadas. En ocasiones, la densidad es muy débil y cuando se encuentra incrustado en el tejido normal de la mama puede ser muy difícil de percibir [36].

Para obtener la textura lineal se aplica el procedimiento propuesto por [33]. Dada la magnitud $G(x, y)$ y la fase o la orientación $\Phi(x, y)$, se construye la suma de vectores de doble ángulo, que se representa mediante la siguiente ecuación:

$$z = Ce^{2\theta} \equiv C * cos(2\theta) + C * sen(2\theta) \tag{3.33}$$

donde $C = G(x, y)$ y $\theta = \Phi(x, y)$.

Para obtener el descriptor de textura lineal, primeramente se calcula la longitud total de cada uno de los componentes de $G(x, y)$ y $\Phi(x, y)$ mediante la siguiente ecuación:

$$z_1 = \sqrt{(\sum C * cos(2\theta))^2 + (\sum C * cos(2\theta))^2} \tag{3.34}$$

Posteriormente, se obtiene la longitud total de todos los vectores mediante la siguiente ecuación:

$$z_2 = \sum \sqrt{(C * cos(2\theta))^2 + (C * cos(2\theta))^2} \tag{3.35}$$

Finalmente, se obtiene el descriptor de textura lineal a través de la siguiente ecuación:

$$TL = \frac{z_1}{z_2} \tag{3.36}$$

3.3. Descriptores de textura

La textura es una de las características más importantes para la identificación de objetos de interés en una imagen [31].

Existen diversos métodos para obtener la textura de una imagen, tales como, la Matriz de Co-ocurrencia de Niveles de Gris (*Gray Level Co-occurrence Matrix* - GLCM), Matriz de Diferencias de Niveles de Gris (*Gray Level Difference Method* - GLDM) y Matriz de Longitud de Secuencias de Niveles de Gris (*Gray Level Run Length Matrix* - GLRLM), entre otras técnicas.

3.3.1. Matriz de Co-ocurrencia de Niveles de Gris (GLCM)

Este método fue publicado en 1973 por Robert M. Haralick [31]. Para construir la matriz GLCM, se define un operador de posición P y una matriz A de tamaño LxL (L es el número de niveles de gris). El elemento a_{ij} es el número de veces que aparecen los píxeles con el nivel de gris z_i (en la posición especificada por P), dado un ángulo θ y una distancia d definida por el incremento Δx y Δy en la imagen original (Figura 3.4).

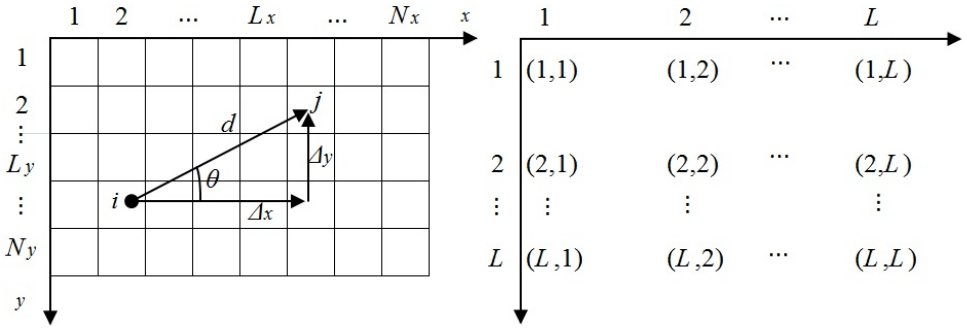


Figura 3.4: a) Imagen original y b) Matriz GLCM.

Entonces, la matriz GLCM se define como [30]:

$$C = \frac{1}{n}A \tag{3.37}$$

Donde n representa el número total de pares de puntos que satisfacen el operador de posición P. De tal manera, que $c_{i,j}$ es una estimación de

la probabilidad conjunta de que un par de puntos que satisfacen P, tenga valores (z_i, z_j).

A continuación se muestra un ejemplo del cálculo de la matriz GLCM con una distancia $d = 1$ y el ángulo $\theta = 0°$.

$$I(x,y) = \begin{bmatrix} 1 & 1 & 2 & 2 & 2 \\ 1 & 1 & 2 & 2 & 2 \\ 1 & 3 & 3 & 3 & 3 \\ 3 & 3 & 4 & 4 & 4 \\ 3 & 3 & 4 & 4 & 4 \end{bmatrix} A = \begin{bmatrix} 2 & 2 & 1 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 5 & 2 \\ 0 & 0 & 0 & 4 \end{bmatrix} C = \frac{1}{20} \begin{bmatrix} 2 & 2 & 1 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 5 & 2 \\ 0 & 0 & 0 & 4 \end{bmatrix}$$

Los descriptores de textura que se extraen a partir de la matriz GLCM se muestran a continuación [51]:

Momento Angular de Segundo Orden. Cuantifica la uniformidad de la textura, tomando valores altos en regiones homogéneas y bajos en regiones que no lo son. Esta medida se calcula de la siguiente manera:

$$SMA = \sum_{i=1}^{Ng} \sum_{j=1}^{Ng} \big(p(i,j)\big)^2 \tag{3.38}$$

donde Ng es el número de niveles de gris en la imagen.

Contraste. Es un estadístico que permite medir transiciones fuertes o variaciones bruscas de niveles de intensidad en la imagen. Esta medida se obtiene mediante la siguiente fórmula:

$$contraste = \sum_{n=0}^{Ng-1} i^2 \Big\{ \sum_{i=1}^{Ng} \sum_{j=1}^{Ng} p(i,j), |i-j| = n \Big\} \tag{3.39}$$

Entropía. Es una medida que registra la aleatoriedad de las intensidades de grises, tomando valores más altos en regiones más homogéneas. Este estadístico se obtiene mediante la siguiente ecuación:

$$entropia = -\sum_{i=1}^{Ng} \sum_{j=1}^{Ng} p(i,j) log\{p(i,j)\} \tag{3.40}$$

Energía. Mide la homogeneidad de una imagen.

$$energia = \sqrt{\sum_{i=1}^{Ng}\sum_{j=1}^{Ng}(p(i,j))^2} \tag{3.41}$$

Correlación. Mide la dependencia lineal de los niveles de gris entre los píxeles y posiciones específicas relacionadas con cada uno de ellos.

$$correlacion = \sum_{i=1}^{Ng}\sum_{j=1}^{Ng}\frac{(i-\mu_x)(i-\mu_y)p(i,j)}{\sigma_x\sigma_y} \tag{3.42}$$

donde μ_x, μ_y, σ_x y σ_y son las medias y desviaciones estándar de p_x y p_y. Cada uno de estos valores se obtienen de la siguiente manera.

$$\mu_x = \sum_{i=1}^{Ng} ip_x \qquad y \qquad \mu_y = \sum_{j=1}^{Ng} ip_y \tag{3.43}$$

$$\sigma_x = \sum_{i=1}^{Ng}(i-\mu_x)^2p_x \qquad y \qquad \sigma_y = \sum_{j=1}^{Ng}(j-\mu_y)^2p_y \tag{3.44}$$

$$p_x(i) = \sum_{j=1}^{Ng} p(i,j) \qquad y \qquad p_y(j) = \sum_{i=1}^{Ng} p(i,j) \tag{3.45}$$

Momento de la Diferencia Inversa. Proporciona la variación de la distribución de niveles de gris en la imagen.

$$MDI = \sum_{i=1}^{Ng}\sum_{j=1}^{Ng}\frac{1}{1+(i-j)^2}p(i,j) \tag{3.46}$$

Varianza. Describe la definición y agrupación de los elementos de la matriz de co-ocurrencia. Este valor aumenta cuando la distancia entre los elementos de la matriz con respecto a la diagonal principal es baja.

$$varianza = \sum_{i=1}^{Ng}\sum_{j=1}^{Ng}(i-\mu)^2 p(i,j) \tag{3.47}$$

Suma Promedio. Mide el promedio de niveles de gris de la región.

$$SP = \sum_{i=2}^{2Ng} i p_{x+y}(i) \tag{3.48}$$

donde

$$p_{x+y}(k) = \sum_{i=1}^{Ng}\sum_{j=1}^{Ng} p(i,j) : i+j=k, k=2,3,...,2Ng \tag{3.49}$$

Suma Varianza. Mide la disimilitud de la distribución de niveles de gris de la región.

$$SV = \sum_{i=2}^{2Ng}(i-SE)^2 p_{x+y}(i) \tag{3.50}$$

Suma Entropía. Mide la acumulación de la aleatoriedad de las intensidades de niveles de gris de la región.

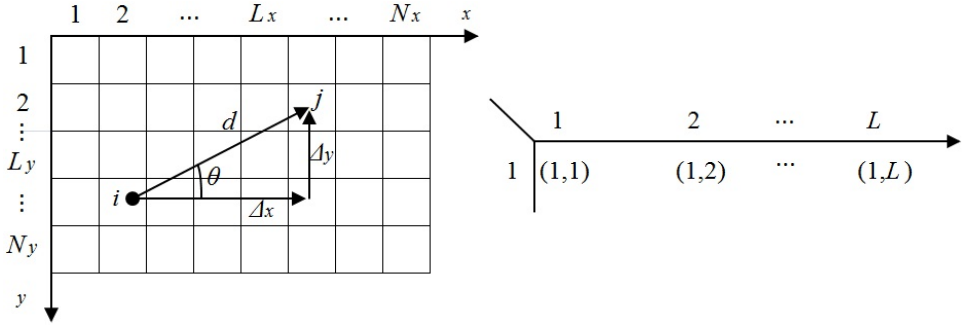
$$SE = -\sum_{i=2}^{2Ng} p_{x+y}(i) log\{p_{x+y}(i)\} \tag{3.51}$$

Homogeneidad. Este valor es grande si todos los valores que están en la diagonal principal son grandes.

$$MDI = \sum_{i=1}^{Ng}\sum_{j=1}^{Ng}\frac{1}{1+(i-j)} p(i,j) \tag{3.52}$$

3.3.2. Matriz de Diferencias de Niveles de Gris (GLDM)

El método GLDM es una técnica de análisis de textura basada en la diferencia absoluta entre pares de niveles de gris o el promedio del nivel de gris de una imagen. Con este método se obtiene un vector H(θ, d), de tamaño igual al número de niveles de gris de la imagen, donde d es la distancia entre el par de píxel y θ es la dirección (Figura 3.5) [2].



(a) (b)

Figura 3.5: a) Imagen original y b) Matriz GLDM.

Dada una función de la intensidad de la imagen $I(i,j)$ y un vector de desplazamiento δ=(Δx,Δy), la diferencia absoluta se obtiene de la siguiente manera:

$$I_\delta(i,j) = |I(i,j) - I(i+\Delta x, j+\Delta y)| \tag{3.53}$$

Y la densidad de probabilidad de $I_\delta(i,j)$ se denota como p_δ.

A continuación se muestra un ejemplo del cálculo de la matriz GLDM con una distancia $d = 1$ y el ángulo $\theta = 0°$.

$$I(x,y) = \begin{bmatrix} 1 & 1 & 2 & 2 & 2 \\ 1 & 1 & 2 & 2 & 2 \\ 1 & 3 & 3 & 3 & 3 \\ 3 & 3 & 4 & 4 & 4 \\ 3 & 3 & 4 & 4 & 4 \end{bmatrix} \; GLDM = \begin{bmatrix} 15 & 4 & 1 & 0 & 0 \end{bmatrix}$$

Los descriptores de textura que se obtienen a partir de este método se describen a continuación:

Media. Describe el grosor de la textura. La media es pequeña cuando los valores de $p_\delta(i)$ están concentrados cerca del píxel de referencia y es grande cuando los valores se encuentra lejos del píxel de referencia.

$$MEDIA = \frac{1}{N}\sum_{i=1}^{N} ip_\delta(i) \tag{3.54}$$

Entropía. Mide la homogeneidad del histograma. La entropía es grande cuando los valores de $p_\delta(i)$ son iguales y es pequeña cuando los valores son desiguales.

$$ENT = -\sum_{i=1}^{N} p_\delta(i)log(p_\delta(i)) \tag{3.55}$$

Contraste. Este es el segundo momento de $p_\delta(i)$ es decir, su momento de inercia con respecto al origen.

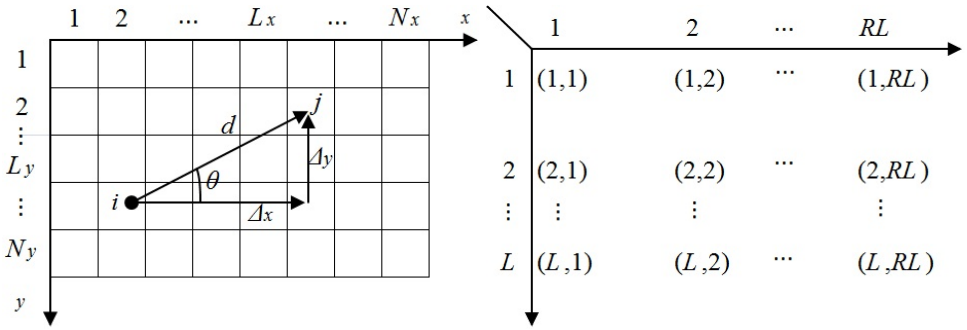
$$CON = \sum_{i=1}^{N} i^2 p_\delta(i) \tag{3.56}$$

Varianza. La varianza es una medida de la dispersión de las diferencias de nivel de gris con respecto a una distancia d.

$$\sigma_d^2 = \sum_{i=1}^{N} (i - MEDIA)^2 p_\delta(i) \tag{3.57}$$

3.3.3. Matriz de Longitud de Secuencias de Niveles de Gris (GLRLM)

La matriz GLRLM codifica la información de textura basado en el número de veces que cada nivel de gris aparece en la imagen por sí misma [66]. Para una imagen determinada, cada elemento $p(i,j|\theta)$ de la matriz de longitud de secuencias representa el número total de secuencias con píxeles de nivel de gris i y con longitud j en una cierta dirección θ (Figura 3.6) [27, 75].



(a) (b)

Figura 3.6: a) Imagen original y b) Matriz GLRLM.

A continuación se muestra un ejemplo del cálculo de la matriz GLRLM con una distancia $d = 1$ y el ángulo $\theta = 0°$.

$$I(x,y) = \begin{bmatrix} 1 & 1 & 2 & 2 & 2 \\ 1 & 1 & 2 & 2 & 2 \\ 1 & 3 & 3 & 3 & 3 \\ 3 & 3 & 4 & 4 & 4 \\ 3 & 3 & 4 & 4 & 4 \end{bmatrix} \; GLRLM = \begin{bmatrix} 1 & 2 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 2 & 0 & 1 \\ 0 & 0 & 2 & 0 \end{bmatrix}$$

Las mediciones que se obtienen a partir de este método son las siguientes:

Énfasis de Secuencia Corta (*Short Run Emphasis - SRE*). Mide la distribución en una longitud de secuencia corta. Es altamente dependiente de la ocurrencia de longitud de secuencias cortas y toma valores grandes para texturas finas.

$$SRE = \frac{\sum_{i=1}^{Ng} \sum_{j=1}^{Nr} \frac{p(i,j)}{j^2}}{\sum_{i=1}^{Ng} \sum_{j=1}^{Nr} p(i,j)} \tag{3.58}$$

Énfasis de Secuencia Larga (*Long Run Emphasis - LRE*). Mide la distribución de longitud de secuencias largas y es altamente dependiente de la aparición de secuencias largas y puede tomar valores grandes para texturas gruesas.

$$LRE = \frac{\sum_{i=1}^{Ng}\sum_{j=1}^{Nr} j^2 p(i,j)}{\sum_{i=1}^{Ng}\sum_{j=1}^{Nr} p(i,j)} \tag{3.59}$$

Desigualdad de los niveles de gris (*Gray Level Non Uniformity - GLNU*). Mide la similitud de los valores de nivel de gris a través de la imagen. Esta característica es pequeña cuando la intensidad de los píxeles son iguales en toda la imagen.

$$GLNU = \frac{\sum_{i=1}^{Ng}\left(\sum_{j=1}^{Nr} p(i,j)\right)^2}{\sum_{i=1}^{Ng}\sum_{j=1}^{Nr} p(i,j)} \tag{3.60}$$

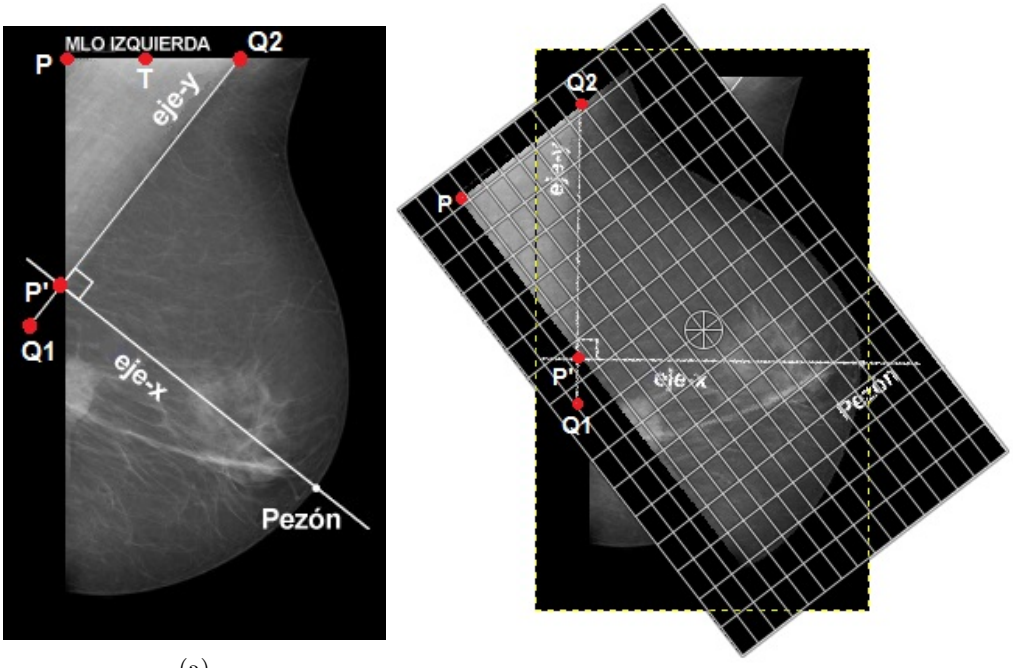
Longitud de Secuencias no Uniforme (*Run Length Non Uniformity - RLNU*). Mide la similitud de la longitud de secuencias a través de la imagen. Esta característica puede ser pequeña cuando la longitud de secuencias son iguales en toda la región.

$$RLNU = \frac{\sum_{i=1}^{Nr}\left(\sum_{j=1}^{Ng} p(i,j)\right)^2}{\sum_{i=1}^{Ng}\sum_{j=1}^{Nr} p(i,j)} \tag{3.61}$$

3.4. Ubicación relativa

La ubicación relativa de una lesión es importante debido a que la mayoría de las masas malignas (45 %) se desarrollan en el cuadrante superior externo de la mama. Para obtener las características de la ubicación relativa sobre

mamografías MLO, es necesario considerar un nuevo sistema de coordenadas que considere el eje que forma el músculo pectoral. La nueva coordenada toma el eje y como el borde del músculo pectoral y el eje x se determina trazando una línea perpendicular a partir del eje y hacia el punto máximo del borde de la mama. Se supone que al final de esta línea se encuentra el pezón de la mama (Figura 3.7) [62].



(a) (b)

Figura 3.7: a) Ejes x y y para el cálculo de la ubicación. b) Cambio de coordenadas de la ubicación de la masa [62].

Para realizar el cambio de coordenadas, primero, se traslada el origen de la imagen (punto P) hacia el punto T, donde T es el punto medio de la distancia de P a $Q2$ (Figura 3.7a). Donde P y $Q2$ son los puntos, $P = (p_1, p_2)$ y $Q2 = (q2_1, q2_2)$. El nuevo origen $T = (id, jd)$, de la imagen se obtiene al aplicarle la siguiente transformación [53].

$$x = i + i_d \tag{3.62}$$

$$y = j + j_d \tag{3.63}$$

Después de conseguir la traslación, la imagen se debe rotar un ángulo θ que corresponde a la inclinación del músculo pectoral. La rotación se consigue mediante las siguientes ecuaciones [53].

$$x = cos\theta i - sen\theta j \tag{3.64}$$

$$y = sen\theta i + cos\theta j \tag{3.65}$$

Y el ángulo θ se obtiene a partir de la siguiente ecuación,

$$cos\theta = \frac{|(p_1 * q2_1) + (p_2 * q2_2)|}{\sqrt{p_1^2 + p_2^2} * \sqrt{q2_1^2 + q2_2^2}} \tag{3.66}$$

Con este nuevo sistema de coordenada se obtiene la ubicación del centro de la lesión (C_x, C_y), considerando el punto P' como el origen del centro de coordenadas como se muestra en la Figura 3.7b. Estos valores de ubicación se normalizan mediante un radio efectivo de la mama, como se muestra en la siguiente ecuación,

$$r = \sqrt{\frac{A}{\pi}} \tag{3.67}$$

Donde A es el área de la mama. Por lo tanto, los valores de la ubicación del centro de la lesión (C_x, C_y), se normalizan de la siguiente manera,

$$U_x = \frac{C_x}{r} \tag{3.68}$$

$$U_y = \frac{C_y}{r} \tag{3.69}$$

Capítulo 4

Redes Bayesianas

En este capítulo se presentan algunos conceptos básicos relacionados con las redes Bayesianas. También, se describen los modelos Naïve Bayes, Naïve Bayes aumentado a árbol, redes Bayesianas K-Dependiente y Naïve Bayes aumentado a bosque, que son utilizados como clasificadores.

4.1. Definición de red Bayesiana

Una red Bayesiana es un modelo gráfico acíclico dirigido, compuesto por un conjunto de nodos que representan variables aleatorias $\{X_1, X_2, ..., X_n\}$ y los arcos que representan relaciones de dependencia directa [21]. Las redes Bayesianas son consideradas una representación compacta, intuitiva y robusta. Pueden representar de manera simultánea la dimensión cualitativa y cuantitativa de un problema. Una de sus grandes ventajas es que pueden trabajar con datos incompletos. También, permiten reducir el sobreajuste de los datos y permiten combinar el conocimiento previo que se tiene con respecto al dominio del problema con datos experimentales [43].

Una definición formal para las redes Bayesianas es la siguiente:

Una red Bayesiana es un par (D, P)*, donde* D *es un grafo dirigido acíclico,* $P = \{p(x_1|\pi_1), ..., p(x_n|\pi_n)\}$ *es un conjunto de* n *funciones de probabilidad condicionada, una para cada variable, y* π_i *es el conjunto de padres del nodo* X_i *en* D*. El conjunto* P *define una función de probabilidad asociada mediante la factorización* [9],

$$p(x) = \prod_{i=1}^{n} p(x_i|\pi_i) \tag{4.1}$$

4.2. Aprendizaje de redes Bayesianas

Construir una red Bayesiana a partir de un conjunto de datos implica un proceso de aprendizaje que se divide en dos partes: estructural y paramétrico. El aprendizaje estructural consiste en obtener la estructura de la red Bayesiana, es decir las relaciones de dependencia e independencia condicional en las variables involucradas. El aprendizaje paramétrico se refiere a la obtención de las probabilidades asociadas dada la estructura de la red [9].

4.2.1. Aprendizaje estructural

En esta fase del proceso de aprendizaje se debe encontrar las relaciones cualitativas entre las variables involucradas. Sin embargo, el problema de encontrar la estructura exacta es imposible, debido a que el problema del aprendizaje es un problema NP-completo. Normalmente, se aplican restricciones en el tipo de estructura con la finalidad de reducir el espacio de búsqueda. Por lo tanto, es necesario encontrar un algoritmo que genere estructuras aproximadamente óptimas y una función de evaluación que proporcione para cada estructura generada, lo cerca que está de representar al conjunto de datos observados [48, 28].

4.2.2. Aprendizaje paramétrico

Esta fase consiste en la estimación de los parámetros numéricos del modelo gráfico probabilístico. Se parte de una estructura ya conocida y de un conjunto de datos asociados. Cuando se tienen datos completos y suficientes para todas las variables en el modelo, es relativamente fácil obtener los parámetros, asumiendo que la estructura está dada. Uno de los métodos más comunes, es el estimador de máxima verosimilitud, en el cual se estiman las probabilidades con base a la frecuencia de los datos [48].

4.3. Redes Bayesianas como Clasificadores

4.3.1. Naïve Bayes

Es uno de los modelos probabilísticos más simples pero muy usados en tareas de clasificación. A pesar de su simplicidad, se muestra competitivo ante otros métodos más sofisticados en diversos ámbitos específicos. Naïve Bayes es el modelo de clasificación construido bajo la premisa de que todas las variables predictoras o atributo son condicionalmente independientes de la variable de clase C [49].

El método de clasificación Naïve Bayes se basa en el teorema de Bayes para predecir para cada instancia x la clase $c \in C$ con la máxima probabilidad a posteriori, como se muestra en la siguiente ecuación [49]:

$$p(c|x) \quad \alpha \quad P(c,x) = \prod_{i=1}^{n} p(x_i|c), \tag{4.2}$$

donde $p(x_i|c)$ representa la probabilidad condicional de $X_i = x_i$ dado que $C = c$. Como resultado el clasificador Naïve Bayes predice en base a [49]:

$$c^* = arg\max_{c} p(c) \prod_{i=1}^{n} p(x_i|c) \tag{4.3}$$

En la Figura 4.1 se muestra la topología típica de un clasificador Naïve Bayes.

El modelo Naïve Bayes ha ido comprobando su potencialidad y robustez en problemas de clasificación supervisada [22, 41, 35]. Este modelo puede hacer predicciones a partir de datos parciales y además su proceso de ejecución es muy rápido. Y entre sus principales desventajas, es de no ser apto para el manejo de variables aleatorias continuas [41].

4.3.2. Naïve Bayes aumentado a árbol (*Tree Augmented Naïve Bayes - TAN*)

Es una red Bayesiana que permite dependencias entre las variables predictoras, donde el conjunto de los nodos padre de la variable a clasificar C,

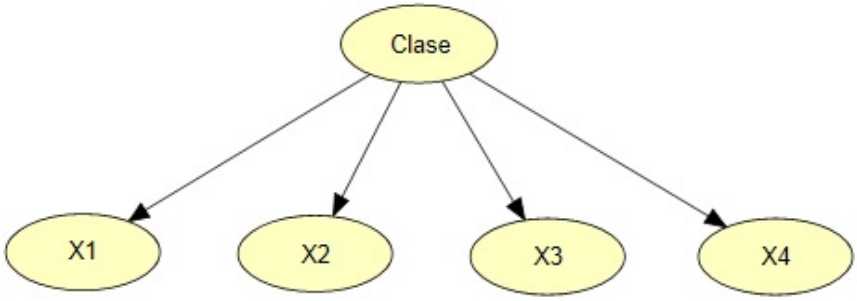


Figura 4.1: Estructura del clasificador Naïve Bayes.

es vacío, mientras que el conjunto de nodos padre de cada una de las variables predictoras (o variables atributo), contienen necesariamente la variable a clasificar, y a lo más otra variable. En la Figura 4.2 se muestra un ejemplo de la topología de una red Bayesiana TAN.

Para construir la estructura de la red Bayesiana TAN se necesita previamente aprender las dependencias entre las diferentes variables predictoras $X_1, ..., X_n$ [49]. Friedman [26] propone un método basado en el algoritmo de Chow-Liu [14], en donde utiliza una función para medir la información que la variable Y proporciona sobre la variable X, cuando el valor C es conocido, como se muestra en la ecuación 4.4. Este algoritmo garantiza que la estructura TAN obtenida tiene asociada la máxima verosimilitud entre todas las posibles estructuras de TAN [39, 58].

$$I_P(X, Y|C) = \sum_{x,y,c} P(x, y, c) log \frac{P(x, y, c)}{P(x|c)P(y|c)} \tag{4.4}$$

Para mayores detalles del algoritmo TAN ver sección B.1.

4.3.3. Clasificador Bayesiano K-Dependiente (*K-Dependence Bayesian Classifier - KDB*)

Este clasificador tiene como objetivo mejorar al clasificador TAN con respecto a las restricciones de que los atributos sean condicionalmente independientes entre sí, ya que supera la restricción del máximo de dos padres de cada variable predictora. Un clasificador Bayesiano K-Dependiente, es una red Bayesiana que contiene la estructura de un clasificador Naïve Bayes

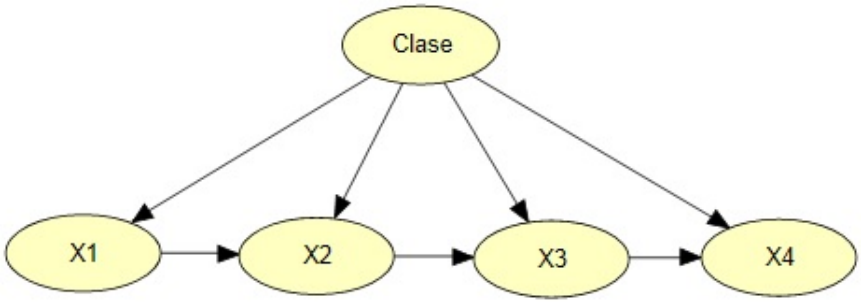


Figura 4.2: Estructura del clasificador TAN.

y permite a cada atributo X_i tener un máximo de k atributos como nodos padre. Descrito de otra manera, $P_a(X_i = \{C, X_{pai}\})$, donde X_{pai} es un conjunto compuesto como máximo de k atributos, y $P_a(C) = 0$ [60, 28]. De tal manera que el modelo Naïve Bayes se corresponde con un clasificador Bayesiano 0-Dependiente, el modelo TAN sería un clasificador Bayesiano 1-Dependiente [24]. En la Figura 4.3 se muestra un ejemplo de la estructura de un clasificador Bayesiano K-Dependiente.

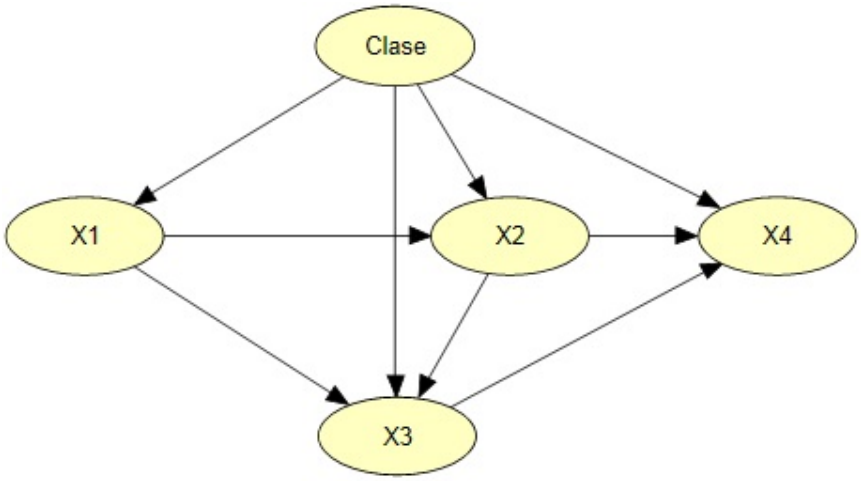


Figura 4.3: Estructura del clasificador Bayesiano K-Dependiente.

Sahami propone en [60] un algoritmo para construir clasificadores Bayesianos K-Dependientes, el cual presenta las siguientes características:

- No obliga incluir dependencias que no existen cuando el valor k es

demasiado grande.

- Es muy adecuado para los dominios de minería de datos debido a que su complejidad computacional es relativamente pequeño.

Una posible desventaja de este algoritmo es que a medida que el valor de k crece, se debe estimar un espacio de probabilidad más grande con la misma cantidad de datos. Esto puede causar estimaciones de probabilidad más inexactas y conduce a una disminución general de la exactitud de predicción.

Para mayor información del algoritmo del clasificador Bayesiano K-Dependiente ver sección B.2.

4.3.4. Naïve Bayes aumentado a bosque (*Forest Augmented Naïve Bayes - FAN*)

Una red Bayesiana FAN permite formar un bosque de árboles disjuntos entre los atributos (ver Figura 4.4). Una red Bayesiana FAN es una mejora de la red Bayesiana TAN. En un modelo TAN, los arcos del árbol formado entre las variables predictoras pueden introducir ruido en la clasificación. La construcción de cada árbol dentro del bosque se realiza de forma similar al modelo TAN, es decir, seleccionando una variable raíz de forma aleatoria y dirigiendo la arista hasta visitar todos los nodos [44].

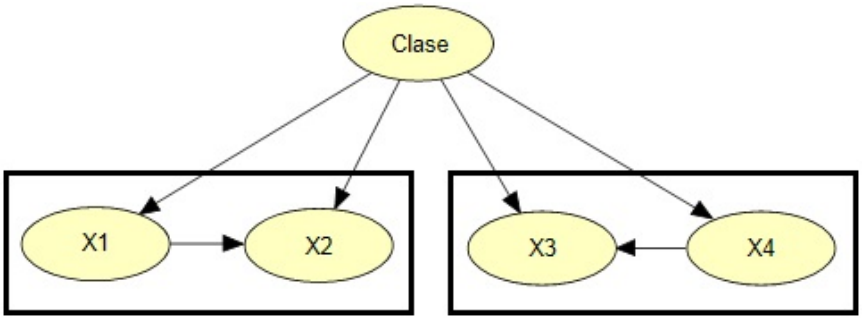


Figura 4.4: Estructura de una red Bayesiana FAN.

Para mayores detalles del algoritmo de red Bayesiana FAN ver sección B.3.

Capítulo 5

Resultados

En este capítulo se presentan los experimentos realizados con los modelos de redes Bayesianas Naïve Bayes, TAN, KDB y FAN, en la clasificación de masas como malignas o benignas. Primero se describe el conjunto de datos y las métricas utilizadas para estimar el desempeño de los modelos. Posteriormente, se presentan los resultados obtenidos al utilizar el conjunto completo de características, así como también una selección de éstas utilizando el método de Análisis Discriminante de Fisher de selección de características.

5.1. Hardware y Software utilizado

El aprendizaje de los clasificadores descritos en este proyecto de tesis, fueron implementados en Matlab Versión 7.10.0.499 (R2010a) ®, con la ayuda de los paquetes Bayesian Network Toolbox (BNT) [50] y Structure Learning Package (SLP) [42] para los modelos de redes Bayesianas; y la biblioteca LIBSVM para la SVM [10]. Otro programa que se utilizó fue ImageJ (Ver. 1.45s) para extraer regiones de interés (masa) en imágenes de mamografía [52]. Y en el lenguaje de programación Java (IDE Eclipse Ver. 4.2.1), fue implementado el sistema de clasificación de masas propuesto en este proyecto de tesis. La ejecución de pruebas se realizó en una estación de trabajo (Workstation) con un procesador Intel(R) Xeon(R) de 2.00 GHz, con Sistema Operativo Windows 7 de 64 bits y memoria RAM de 16 GB.

5.2. Conjunto de datos y configuración experimental

El conjunto de imágenes de mamografías que se ocupa para este proyecto de tesis se tomaron de la base de datos mini-MIAS [67]. Esta es una base de datos pública que proporciona un total de 322 imágenes de la mama derecha e izquierda de cada paciente, y del tipo Media-Lateral-Oblicua con la distribución que se muestra en el Cuadro 5.1. Estas imágenes tienen una resolución de 1024 x 1024 píxeles y cuentan con una profundidad de 8 bits, donde cada píxel registra un tamaño de 0.2 x 0.2 mm del objeto sensado.

Tipo de hallazgos	**# Imágenes**
Normal	206
Masas	59
Calcificaciones	23
Distorsión de la arquitectura	19
Asimetría	15

Cuadro 5.1: Imágenes de la base de datos mini-MIAS.

Para construir el conjunto de datos de estudio de la presente tesis, se seleccionaron las 58 imágenes que presentaban información de masas (ver Cuadro 5.1). La descripción de estas imágenes se muestra en el Cuadro 5.2.

De cada una de estas imágenes, se extrajeron las regiones correspondientes a las masas (regiones de interés), como se muestra en la Figura 5.1. Esta extracción se realizó con la ayuda del programa ImageJ (Ver. 1.45s) [52] y utilizando la información de ubicación y tamaño de la masa contenida en la base de datos.

Clase	**Benigna**	**Maligna**
Circunscritas	20	4
Mal definidas	7	8
Espiculadas	11	8

Cuadro 5.2: Imágenes de mini-MIAS con masa.

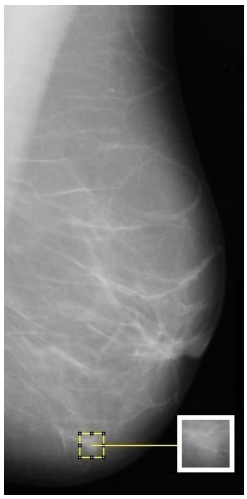


Figura 5.1: Extracción de la región de interés que corresponde a una masa.

Posteriormente, para cada una de las regiones de interés, se calcularon las 67 características que se explican en el Capítulo 3. Para calcular los descriptores de intensidad y textura, previamente se ecualizó y cuantizó la imagen a 7 bits con el objetivo de resaltar la información contenida en la región correspondiente (ver Anexo A). El cálculo de cada característica de textura se realizó para $\theta = \{0°, 45°, 90°$ y $135°\}$ y una distancia $d = 1$, y después, se estimó el promedio y el rango, obteniendo así, 38 descriptores de textura.

Los valores de las características que se obtuvieron son datos continuos y con rangos dinámicos diferentes. Para evitar rangos dinámicos diferentes, los datos fueron normalizados, consiguiendo una media igual a cero y una desviación estándar igual a uno. Después se aplicó el método de Discretización de Igual Anchura con una partición de $k = 2$. Estos métodos de normalización y discretización se describen en el Anexo A.

Los objetivos a mostrar en nuestros experimentos son los siguientes. Primero, evaluar el desempeño de diferentes métodos de redes Bayesianas mediante distintas métricas, en el problema de clasificación de masas. Segundo, encontrar el mejor método de red Bayesiana para dicho problema. Y por último, encontrar la mejor arquitectura (estructura) de la red de manera algorítmica. Cabe mencionar que en la evaluación de los distintos métodos queda implícito el uso de las características extraídas de las regiones de

interés.

5.3. Métricas de desempeño del clasificador

A fin de evaluar el desempeño de los clasificadores, los resultados fueron catalogados como:

- Positivo Verdadero (PV): Cuando la lesión es maligna y el clasificador determina que es maligna.
- Positivo Falso (PF): Cuando el clasificador resuelve que la lesión es maligna, pero la lesión es benigna.
- Negativo Verdadero (NV): Cuando la lesión es benigna y el clasificador determina que es benigna.
- Negativo Falso (NF): Cuando el clasificador resuelve que la lesión es benigna, pero la lesión es maligna.

A partir de los cuatro valores anteriores se obtuvieron las siguientes métricas para evaluar el desempeño de los clasificadores:

- Exactitud (*Accuracy*): Mide la proporción de los resultados verdaderos (PV y NV) del conjunto total de lesiones.

$$exactitud = \frac{Num.\quad de\quad PV + Num.\quad de\quad NV}{Num.\quad total\quad de\quad masas} \tag{5.1}$$

- Sensibilidad (*Sensitivity*): Es un parámetro que mide la probabilidad de clasificar correctamente una lesión maligna. Es el cociente entre positivos verdaderos y el total de lesiones. La sensibilidad varía de 0 a 1 (0 a 100 %). Cuando el valor numérico de esta métrica es alta, entonces, hay mejor capacidad de detectar lesiones malignas.

$$sensibilidad = \frac{Num.\quad de\quad PV}{Num.\quad de\quad PV + Num.\quad de\quad NF} \tag{5.2}$$

- Especificidad (*Specificity*): Es un parámetro que mide la probabilidad de clasificar correctamente a una lesión benigna. Es el cociente entre negativos verdaderos y el total de lesiones benignas. La especificidad varía de 0 a 1 (0 100 %). Cuando el valor numérico de esta métrica es alta, entonces, hay mejor capacidad de detectar lesiones benignas.

$$especificidad = \frac{Num. \quad de \quad NV}{Num. \quad de \quad NV + Num. \quad de \quad PF} \tag{5.3}$$

5.4. Evaluación de redes Bayesianas con el conjunto completo de características

En este experimento se evalúa el desempeño en la clasificación de masas de los modelos de redes Bayesianas Naïve Bayes, TAN, KDB y una FAN, utilizando el conjunto completo de 67 características (ver Cuadro 5.3). Las redes Bayesianas se construyeron y evaluaron en Matlab Versión 7.10.0.499 (R2010a) ®, con la ayuda de los paquetes *Bayesian Network Toolbox BNT* [50] y *Structure Learning Package SLP* [42]. El rendimiento de las cuatro redes Bayesianas fue estimado mediante la técnica de validación cruzada dejando uno fuera (*Leave One Out Cross Validation LOO-CV*).

Los resultados para este experimento se presentan en el Cuadro 5.4. Se puede observar en este cuadro que el desempeño de las redes Bayesianas es muy bajo. El modelo más simple, Naïve Bayes, proporciona una exactitud mayor a las demás, debido a que no tiene relaciones de dependencia. Se observa también, que las redes Bayesianas TAN y FAN proporcionan el menor rendimiento, esto sucede debido al aumento de relaciones de dependencia. Este experimento muestra que probablemente algunas características no son relevantes para la clasificación.

También en el Cuadro 5.4, se presentan, a manera de comparación, los resultados obtenidos por un clasificador denominado Máquina de Soporte Vectorial (*Support Vector Machine - SVM*) [73], el cual es reconocido en el área de aprendizaje automático como uno de los clasificadores más competitivos. Para este fin, hemos utilizado un clasificador SVM implementado en la biblioteca LIBSVM [10]. Se puede apreciar, que aunque el desempeño de la SVM es superior a los modelos de redes Bayesianas en exactitud y especificidad, su capacidad para clasificar masas malignas también es baja.

<table>
<tr><th>Intensidad</th><th colspan="2">Forma</th><th colspan="2">Textura</th><th>Ubicación relativa</th></tr>
<tr><td>Media</td><td colspan="2">Area</td><td rowspan="15"></td><td>Momento Angular de Segundo Orden</td><td>U_x</td></tr>
<tr><td>Mediana</td><td colspan="2" rowspan="2">Perímetro</td><td>Contraste</td><td>U_y</td></tr>
<tr><td>Varianza</td><td rowspan="14"></td><td rowspan="2">Entropía</td></tr>
<tr><td>Asimetría</td><td colspan="2">Compacidad</td></tr>
<tr><td>Kurtosis</td><td rowspan="5">Longitud radial normalizada</td><td>Media</td><td rowspan="2">Energía</td></tr>
<tr><td rowspan="2">Contraste</td><td>Desviación estándar</td></tr>
<tr><td rowspan="4"></td><td>Entropía</td><td rowspan="2">GLCM</td><td>Correlación</td></tr>
<tr><td>Índice de área</td><td>Momento de la diferencia inversa</td></tr>
<tr><td>Rugosidad del contorno</td><td>Varianza</td></tr>
<tr><td>Cruce por cero</td><td>Suma promedio</td></tr>
<tr><td rowspan="7">Momentos invariantes de Hu</td><td>Φ_1</td><td>Suma varianza</td></tr>
<tr><td>Φ_2</td><td>Suma Entropía</td></tr>
<tr><td rowspan="2"></td><td>Homogeneidad</td></tr>
<tr><td>Φ_3</td><td rowspan="4">GLDM</td><td rowspan="2">Media</td></tr>
<tr><td>Φ_4</td></tr>
<tr><td>Φ_5</td><td rowspan="3">Entropía</td></tr>
<tr><td>Φ_6</td></tr>
<tr><td>Φ_7</td><td>Contraste</td></tr>
<tr><td rowspan="4">Patrónes estrellados</td><td>f_1</td><td>varianza</td></tr>
<tr><td>f_1 promedio</td><td rowspan="4">GLRLM</td><td rowspan="2">Énfasis de Secuencia Corta</td></tr>
<tr><td rowspan="2">f_2</td></tr>
<tr><td>f_2 promedio</td><td>Énfasis de Secuencia Larga</td></tr>
<tr><td colspan="2" rowspan="2">Textura Lineal</td><td>Desigualdad de los Niveles de Gris</td></tr>
<tr><td>Longitud de Secuencias no Uniforme</td></tr>
</table>

Cuadro 5.3: Descriptores de intensidad, forma, textura y ubicación.

Clasificador	**Exactitud**	**Sensibilidad**	**Especificidad**
Naïve Bayes	0.43	0.20	0.55
TAN	0.27	0.20	0.31
K-DB	0.32	0.20	0.39
FAN	0.27	0.20	0.31
SVM	0.59	0.10	0.84

Cuadro 5.4: Desempeño obtenido por las redes Bayesianas y un clasificador SVM, con el conjunto completo de características.

5.5. Evaluación de redes Bayesianas con un subconjunto de características

Debido al bajo desempeño de los clasificadores con el conjunto completo de características y al hecho de que es más entendible una estructura (gráfica) de red más compacta (esto es, con menos nodos), se realizó una selección de características. Para el proceso de selección de características, se usó el método de Análisis Discriminante de Fisher [70]. El análisis de discriminación se aplicó por separado sobre cada uno de los tipos de descriptores (intensidad, forma y textura). De este análisis de discriminación, se tomaron las características con mayor valor de discriminación y se realizaron experimentos combinando y eliminando características, hasta finalmente obtener las 11 características que proporcionaban un mayor rendimiento para las redes Bayesianas. El subconjunto de características obtenido se muestran en el Cuadro 5.5.

Con el conjunto de características seleccionadas se construyeron nuevamente clasificadores del tipo Naïve Bayes, TAN, KDB ($K = 2$) y FAN. Los resultados para estos modelos se presentan en el Cuadro 5.6 y las topologías correspondientes en las figuras 5.2, 5.3, 5.4 y 5.5. Se puede apreciar en este cuadro, que la selección de características ayudó a mejorar el desempeño de los clasificadores. En este caso, las redes Bayesianas con mayor rendimiento fueron la TAN y la FAN, mientras que el modelo Naïve Bayes proporciona el menor rendimiento. Estos dos mejores modelos, tienen una capacidad aceptable para clasificar masas benignas, y regular, para masas malignas. El desempeño superior a Naïve Bayes de los modelos TAN, FAN y KDB, demuestran que existen dependencias importantes entre las características, las cuales, considerando los resultados, fueron capturadas mejor por la TAN

Intensidad	**Forma**	**Textura**
F1: Media	F3: Área	F11: Varianza (GLCM-Rango)
F2: Mediana	F4: Perímetro	
	F5: Compacidad	
	F6: Media LRN	
	F7: Desviación Estándar LRN	
	F8: Cruce Cero LRN	
	F9: Momentos Invariantes de Hu3	
	F10: Textura Lineal	

Cuadro 5.5: Subconjunto de características seleccionadas con base al Análisis Discriminante de Fisher.

y FAN. De hecho, como se muestran en las figuras 5.3 y 5.5, estos modelos son casi equivalentes.

En el Cuadro 5.6, también se puede observar que los resultados de las redes Bayesianas son mejores a los de un clasificador SVM. Estos resultados indican que el conocimiento sobre la clasificación de las masas, implícito en los datos, es mejor descrito por las redes Bayesianas.

El desempeño de las redes Bayesianas podría incrementarse al utilizar otro tipo de características que ayuden a describir mejor las irregularidades en forma, márgenes y densidad de las masas.

Clasificador	**Exactitud**	**Sensibilidad**	**Especificidad**
Naïve Bayes	0.63	0.4	0.76
TAN	0.81	0.65	0.89
K-DB	0.74	0.55	0.84
FAN	0.81	0.65	0.89
SVM	0.76	0.60	0.84

Cuadro 5.6: Desempeño obtenido por las redes Bayesianas y un clasificador SVM, con el subconjunto de once características.

Por otra parte, las características consideradas en los modelos de redes Bayesianas nos indican que tamaño, forma, márgenes y densidad de la masa

son parámetros importantes en el diagnóstico. El tamaño se describe con las características F3 y F4; la forma, con F5 y F9; los márgenes, con F6, F7, F8 y F10; y la densidad, con F1, F2 y F11. De estas once características, nueve son consistentes con la literatura médica; sólo el tamaño no es considerado relevante para el diagnóstico, pero es un parámetro importante que ayuda al radiólogo a determinar el tratamiento adecuado [65].

Al analizar las topologías de las redes Bayesianas y considerando los resultados, nos podemos percatar que los clasificadores que incluyen relaciones entre las características modelan mejor la información de las masas. Una interpretación subjetiva, para las relaciones causales contempladas en las topologías de las mejores redes TAN y FAN (que se muestran en las figuras 5.3 y 5.5, respectivamente) sería la siguiente. La característica que describe el tamaño de la masa tiene influencia en la forma, factor que a su vez afecta a los márgenes. También, en estos modelos se puede identificar, que los márgenes repercuten en el tamaño y en la densidad de la masa; y que la densidad influye tanto en los márgenes como en la forma. Aunque es difícil determinar la validez de estas relaciones sin la ayuda de un experto, considerando la literatura médica, se puede decir que estas relaciones reflejan la manera en la que el experto efectúa el análisis de una masa para determinar su diagnóstico [65, 32].

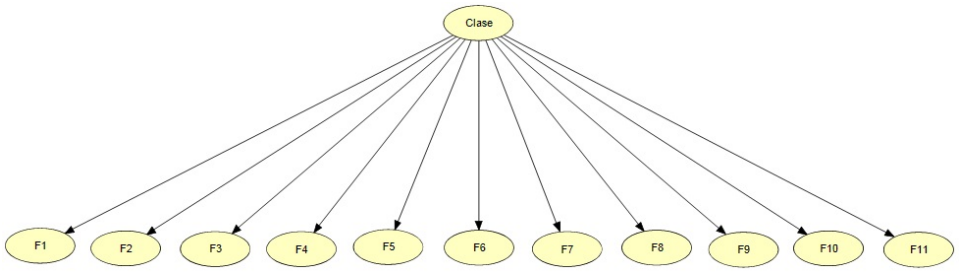


Figura 5.2: Topología de la red Bayesiana Naïve Bayes obtenida con el subconjunto de once características.

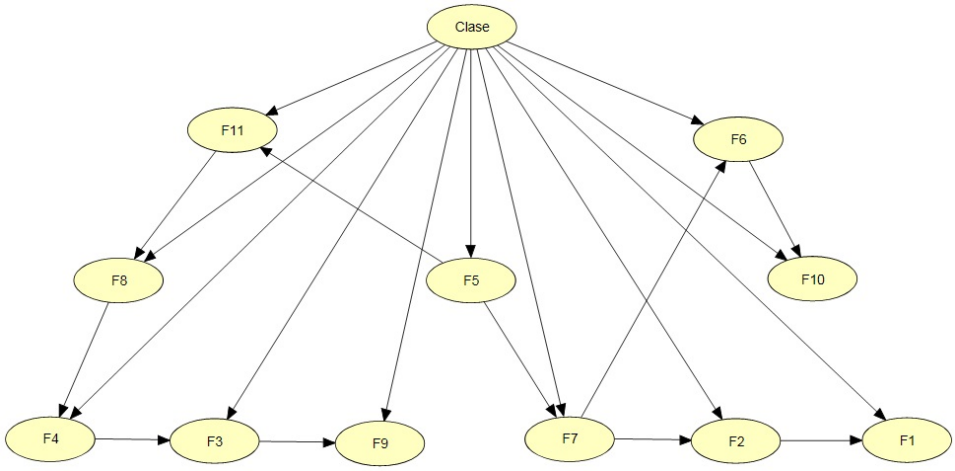


Figura 5.3: Topología de la red Bayesiana TAN obtenida con el subconjunto de once características.

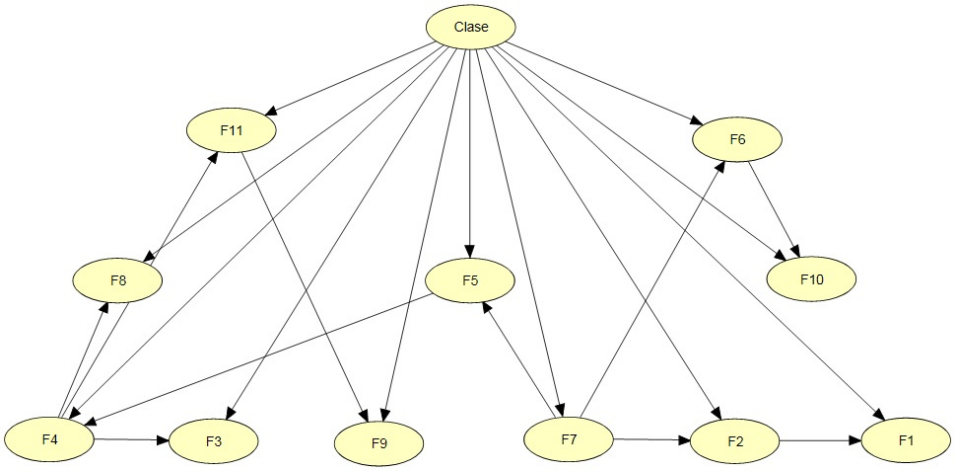


Figura 5.4: Topología de la red Bayesiana KDB obtenida con el subconjunto de once características.

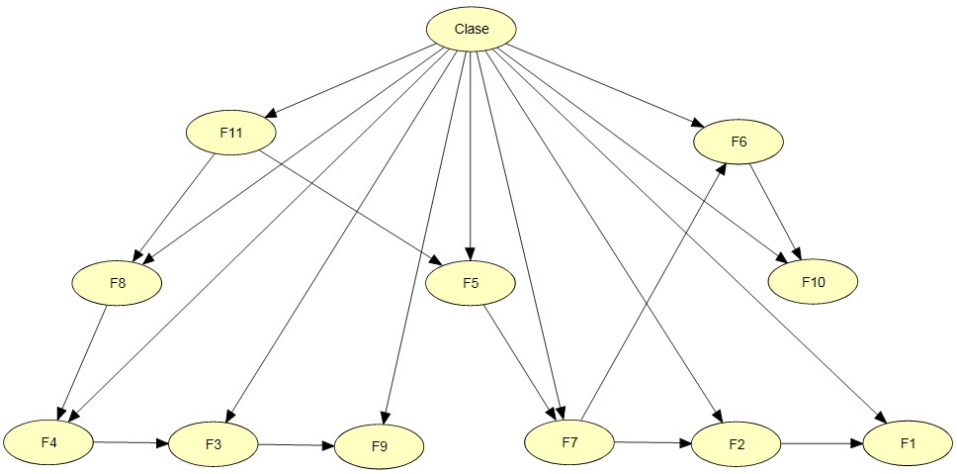


Figura 5.5: Topología de la red Bayesiana FAN obtenida con el subconjunto de once características.

Capítulo 6

Conclusiones y trabajos futuros

En este trabajo se analizó el desempeño de redes Bayesianas del tipo Naïve Bayes, TAN y FAN, en la discriminación de masas benignas y malignas. Los clasificadores fueron entrenados a partir de características extraídas de forma automática de 58 imágenes de masas mamográficas. Para esto, se exploró el uso de diversos tipos de descriptores de objetos en imágenes que ayudaran a obtener las características (forma, márgenes, tamaño y densidad) consideradas por los radiólogos al diagnosticar masas. Los mejores modelos TAN y FAN, obtuvieron un desempeño del 81 % en exactitud. Comparando este resultado con el desempeño promedio del 75 % reportado para los radiólogos [11] y con el 76 % de un clasificador SVM, se puede concluir que es prometedor el uso de redes Bayesianas con características extraídas de forma automática, en la clasificación de masas de mamografías.

Por otra parte, considerando los resultados obtenidos por los mejores modelos de redes Bayesianas del 65 % en sensibilidad y 89 % en especificidad, se puede concluir que el espacio de características analizado es adecuado para describir masas benignas y que es necesario incluir otro tipo de mediciones que ayuden a mejorar la descripción de masas malignas.

Algunas posibles líneas de investigación que ayudarían a enriquecer el trabajo aquí desarrollado, son las siguientes:

- Realizar estudios con un grupo de radiólogos expertos para determinar la validez de los resultados obtenidos.

- Analizar el desempeño de los modelos de redes Bayesianas y de clasificadores SVM utilizando datos balanceados.
- Analizar otras características que proporcionen mejor descripción de masas malignas.
- Incorporar en los modelos de redes Bayesianas información sobre el paciente, como factores de riesgo y datos clínicos.
- Replicar los experimentos aquí realizados con una base de datos que contenga mayor casos de muestras de estudio.
- Realizar comparativas con otros métodos de clasificación.
- Evaluar el desempeño de ensambles de clasificadores.

Bibliografía

[1] Kaushik Adhikary and Amit Kumar. Feature Extraction and Classification Technique in Neural Network. *International Journal of Computer Applications*, 35(3), 2011.

[2] U. Akilandeswari, R. Nithya, and B. Santhi. Review on feature extraction methods in pattern classification. *European Journal of Scientific Research*, 71(2):265–272, 2012.

[3] Corinne Balleyguier, Salma Ayadi, Kim Van Nguyen, Daniel Vanel, Clarisse Dromain, and Robert Sigal. BIRADS classification in mammography. *European Journal of Radiology*, 61(2):192–194, 2007.

[4] Nanhyo Bang and Kyhyun Um. Structural analysis and matching of shape by logical property. In *Structural, Syntactic, and Statistical Pattern Recognition*, pages 521–529. Springer, 2004.

[5] Jelena Bozek, Kresimir Delac, and Mislav Grgic. Computer-aided detection and diagnosis of breast abnormalities in digital mammography. In *ELMAR, 2008. 50th International Symposium*, volume 1, pages 45–52. IEEE, 2008.

[6] Jelena Bozek, Mario Mustra, Kresimir Delac, and Mislav Grgic. A survey of image processing algorithms in digital mammography. In *Recent Advances in Multimedia Signal Processing and Communications*, pages 631–657. Springer, 2009.

[7] Freddie Bray, Peter McCarron, and D Maxwell Parkin. The changing global patterns of female breast cancer incidence and mortality. *Breast Cancer Research*, 6(6):1–5, 2004.

[8] World Wide Breast Cancer. Breast Cancer Statitistics World Wide. http://www.worldwidebreastcancer.com/learn/breast-cancer-statistics-worldwide/, 2010. [Fecha de consulta: 26 de Enero de 2014].

[9] Enrique Castillo, José Manuel Gutiérrez, and Ali S Hadi. Sistemas expertos y modelos de redes probabilísticas. *Academia de Ingeniería*, 1997.

[10] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):1–27, 2011.

[11] HD Cheng, XJ Shi, Rui Min, LM Hu, XP Cai, and HN Du. Approaches for automated detection and classification of masses in mammograms. *Pattern recognition*, 39(4):646–668, 2006.

[12] Ryszard S Choraś. Shape and texture feature extraction for retrieval mammogram in databases. In *Information Technologies in Biomedicine*, pages 121–128. Springer, 2008.

[13] A L Chouhayd. *El cáncer de mama: observación, educación e intervención del farmacéutico comunitario.* PhD thesis, Universidad Cardenal Herrera, España, 2011.

[14] C. Chow and C. Liu. Approximating discrete probability distributions with dependence trees. *Information Theory, IEEE Transactions on*, 14(3):462–467, 1968.

[15] R. Chrisanthar. *Resistance to Chemotherapy in Breast Cancer: Potential role of p21B, p27 and the p53 apoptotic pathway.* PhD thesis, University of Bergen, Norway, 2008.

[16] A. Darwiche. What are Bayesian networks and why are their applications growing across all fields? *Communication of the ACM*, 53(12), 2010.

[17] Instituto Nacional de Estadística y Geografía. Geografía del cáncer femenino, como causa de muerte. http://www.inegi.org.mx/inegi/contenidos/espanol/prensa/Boletines/Boletin/Comunicados/Especiales/2011/Julio/comunica.pdf, 2011. [Fecha de consulta: 22 de Agosto de 2013].

[18] Secretaría de Salud. *Manual Control de Calidad en Mastografía.* México, 2002.

[19] Pasquale Delogu, Maria Evelina Fantacci, Parnian Kasae, and Alessandra Retico. Characterization of mammographic masses using a gradient-

based segmentation algorithm and a neural classifier. *Computers in Biology and Medicine*, 37(10):1479–1491, 2007.

[20] Thomas M. Deserno. *Biomedical Image Processing (Biological and Medical Physics, Biomedical Engineering)*. Springer, 2011.

[21] Francisco Javier Díez. Introducción al razonamiento aproximado. *Universidad de Educación a Distancia*, 1998.

[22] Pedro Domingos and Michael Pazzani. On the optimality of the simple Bayesian classifier under zero-one loss. *Machine learning*, 29(2-3):103–130, 1997.

[23] Sonia Elias, Alvaro Contreras, and Carlos Llanque. Cáncer o carcinoma de mama. *Revista Paceña Medicina Familiar*, 5(7):14–23, 2008.

[24] Enrique Fernández. Análisis de Clasificadores Bayesianos. *Trabajo Final de Especialidad en Ingeniería de Sistemas Expertos. Escuela de Postgrado. Instituto Tecnológico de Buenos Aires*, 2004.

[25] E. A. Fischer, J. Y. Lo, and M. K. Markey. Bayesian Networks of Bi-RADS Descriptors for Breast Lesion Classification. In *Engineering in Medicine and Biology Society, 2004. IEMBS'04. 26th Annual International Conference of the IEEE*, volume 2, pages 3031–3034. IEEE, 2004.

[26] Nir Friedman, Dan Geiger, and Moises Goldszmidt. Bayesian network classifiers. *Machine Learning*, 29(2-3):131–163, 1997.

[27] Guillermo García, Josu Maiora, Arantxa Tapia, and Mariano De Blas. Evaluation of texture for classification of abdominal aortic aneurysm after endovascular repair. *Journal of Digital Imaging*, 25(3):369–376, 2012.

[28] Francisco Javier García Castellano. *Modelos Bayesianos para la clasificación supervisada. Aplicaciones al análisis de datos de expresión genética*. PhD thesis, Universidad de Granada, 2009.

[29] Duván Alberto Gómez Betancur. *Método de detección de distorsiones de la arquitectura de la glándula mamaria a partir de imágenes radiológicas*. PhD thesis, Universidad Nacional de Colombia, Sede Medellín, 2012.

[30] Rafael C. González and Richard E. Woods. *Tratamiento digital de imágenes*. Addison-Wesley Díaz de Santos, 1996.

[31] Robert M. Haralick, Karthikeyan Shanmugam, and I. Dinstein. Textural features for image classification. *IEEE Transactions on Systems, Man and Cybernetics*, SMC-3(6):610–621, 1973.

[32] V. P. Jackson, K. A. Dines, L. W. Bassett, R. H. Gold, and H. E. Reynolds. Diagnostic importance of the radiographic density of noncalcified breast masses: analysis of 91 lesions. *AJR. American journal of roentgenology*, 157(1):25–28, 1991.

[33] B. Johansson. Backprojection of some image symmetries based on a local orientation description. Technical report, Linköping University, 2000.

[34] Charles E. Kahn Jr, Linda M Roberts, Kun Wang, Deb Jenks, and Peter Haddawy. Preliminary Investigation of a Bayesian Network for Mammographic Diagnosis of Breast Cancer. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, page 208. American Medical Informatics Association, 1995.

[35] Bekir Karlik. Hepatitis Disease Diagnosis Using Backpropagation and the Naive Bayes Classifiers. *IBU Journal of Science and Technology*, 1(1), 2012.

[36] Nico Karssemeijer and Guido M. te Brake. Detection of stellate distortions in mammograms. *IEEE Transactions on Medical Imaging*, 15(5):611–619, 1996.

[37] Lisa M. Kinnard, Shih-Chung B. Lo, Paul C. Wang, Matthew T. Freedman, and Mohammed F. Chouikha. Separation of malignant and benign masses using maximum-likelihood modeling and neural networks. In *Medical Imaging 2002*, pages 733–741. International Society for Optics and Photonics, 2002.

[38] F. Knaul, Rafael Lozano, Héctor Arreola-Ornelas, and H. Gómez-Dantés. México: Numeralia del Cáncer de mama. *Fundación Mexicana para la Salud*, 2011.

[39] P. Larrañaga. Aprendizaje Automático de Modelos Gráficos II. Aplicaciones a la Clasificación Supervisada. In Colección CIENCIA Y TÉCNICA, editor, *Sistemas Expertos Probabilísticos*, España, 2008.

[40] G. Lavanya and D. Sudarvizhi Me. Breast tumour detection and classification using Naïve Bayes classifier algorithm. *International Journal of Emerging trends in Engineering and Development*, 3(2), 2012.

[41] Samuel D. Pacheco Leal, Luis Gerardo Díaz Ortiz, and Rodolfo García Flores. El clasificador Naïve Bayes en la extracción de conocimiento de bases de datos. *Ingenierías*, 8(27):1–25, 2005.

[42] Philippe Leray and Olivier Francois. BNT structure learning package: Documentation and experiments. *Machine Learning Research*, 2004.

[43] P. J. López, G. J. García, S. L. Fuentes, and E. I. Fuente-Solana. Las redes Bayesianas como herramientas de modelado en Psicología. *Anales de Psicología*, 23(2), 2007.

[44] Peter Lucas. Restricted Bayesian Network Structure Learning. In *Advances in Bayesian Networks*, volume 146 of *Studies in Fuzziness and Soft Computing*, pages 217–234. Springer, 2004.

[45] Arnau Oliver Malagelada. *Automatic mass segmentation in mammographic images*. PhD thesis, University of Girona, Spain, 2007.

[46] K. F. Marie, G. Nigenda, R. Lozano, O. H. Arreola, A. Langer, and J. Frenk. Breast cancer in Mexico: a pressing priority. *Reproductive Health Matters*, 16(32), 2008.

[47] S. Menna, M. R. Di-Virgilio, P. Burke, A. Frigerio, E. Boglione, G. Ciccarelli, S. Di-Filipo, and L. Garretti. Diagnostic accuracy of commercial system for computerassisted detection (CADx) as an adjunct to interpretation of mammograms. *US National Library of Medicine*, 4(110), 2005.

[48] E. T. Miquelez. *Avances en Algoritmos de Estimación de Distribuciones. Alternativas en el Aprendizaje y Representación de Problemas*. PhD thesis, Universidad del País Vasco, España, 2010.

[49] V. D. A. Morales. *Clasificadores Bayesianos en la Selección Embrionaria en Tratamientos de Reproducción Asistida*. PhD thesis, Universidad del País Vasco, España, 2008.

[50] K. P. Murphy. The Bayes, Net ToolBox for Matlab. 2001.

[51] R. Nithya and B. Santhi. Classification of normal and abnormal patterns in digital mammograms for diagnosis of breast cancer. *International Journal of Computer Applications*, 28(6), 2011.

[52] National Institutes of Health. ImageJ Image Processing and Analysis in Java. http://rsbweb.nih.gov/ij/, 2004. [Fecha de consulta: 24 de Enero de 2014].

[53] Gonzalo Pajares Martinsanz and J. M. de la Cruz García. Visión por computador. Imágenes digitales y aplicaciones, 2002.

[54] D Maxwell Parkin and Leticia MG Fernández. Use of statistics to assess the global burden of breast cancer. *The Breast Journal*, 12(s1):S70–S80, 2006.

[55] Gianfranco Passariello. *Imagénes médicas. Adquisición, Analisis*. Equinoccio, 1999.

[56] P. Pathmanathan. *Simulating the Deformation of the Breast using Nonlinear Elasticity and the Finite Element Method*. PhD thesis, University of Oxford, United Kingdom, 2006.

[57] César Augusto Poveda. Sistema BI-RADS: Descifrando el informe mamográfico. *Repertorio de Medicina y Cirugía*, 19(1):1–18, 2010.

[58] F. V. Robles. *Clasificadores Supervisada basada en Redes Bayesianas. Aplicación en Biología Computacional*. PhD thesis, Universidad Politécnica de Madrid, España, 2003.

[59] Daniel L Rubin, Elizabeth S Burnside, and Ross Shachter. A Bayesian Network to assist mammography interpretation. In *Operations Research and Health Care*, pages 695–720. Springer, 2005.

[60] Mehran Sahami. Learning Limited Dependence Bayesian Classifiers. In *KDD-96: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pages 335–338. AAAI Press, 1996.

[61] Mehul P. Sampat, Mia K Markey, Alan C. Bovik, et al. Computer-aided detection and diagnosis in mammography. *Handbook of image and video processing*, 2(1):1195–1217, 2005.

[62] M. R. M. Samulski. Classification of Breast Lesions in Digital Mammograms. Master's thesis, University Medical Center Nijmegen, Netherlands, 2006.

[63] Jesús Cárdenas Sánchez and Francisco Sandoval Guerrero. Consenso nacional sobre el diagnóstico y tratamiento del cáncer mamario. *Revista Mexicana de Mastología*, 1:13–38, 2011.

[64] M. Shinde. Computer aided diagnosis in digital mammography: Classification of mass and normal tissue. Master's thesis, University of South Florida, United States of America, 2003.

[65] Edward A. Sickles. Breast Masses: Mammographic Evaluation. *Radiology*, 173(2):297–303, 1989.

[66] T. Stavros. *Image Processing and Analysis Methods in Thyroid Ultrasound Imaging*. PhD thesis, University of Patras, Greece, 2007.

[67] John Suckling, J Parker, DR Dance, S Astley, I Hutt, C Boggis, I Ricketts, E Stamatakis, N Cerneaz, Siew-Li Kok, et al. The mammographic image analysis society digital mammogram database. In *Exerpta Medica*, International Congress Series 1069, pages 375–378, 1994.

[68] Jinshan Tang and Xiaoming Liu. Classification of Breast Mass in Mammography with an Improved Level Set Segmentation by Combining Morphological Features and Texture Features. In *Multi Modality State-of-the-Art Medical Image Segmentation and Registration Methodologies*, pages 119–135. Springer, 2011.

[69] G te Brake. *Computer Aided Detection of Masses in Digital Mammograms*. PhD in medical sciences, Radboud University Nijmegen, 2000.

[70] Sergios Theodoridis, Aggelos Pikrakis, Konstantinos Koutroumbas, and Dionisis Cavouras. *Introduction to Pattern Recognition: A Matlab Approach*. Academic Press, 2010.

[71] Tibor Tot. *Breast Cancer: A Lobar Disease*. Springer, 2011.

[72] Ribate Ander Urruticoechea. *Description and Pre-clinical Validation of Dynamic Molecular Determinants of Sensitivity to Aromatase Inhibitors in Breast Cancer/Descripción y Validación Pre-Clínica de Marcadores Moleculares Dinámicos de Sensibilidad a Inhibidores de la Aromatasa en Cáncer de Mama*. PhD thesis, University of Barcelona, Spain, 2007.

[73] Vladimir N. Vapnik. Statistical Learning Theory. *Wiley-Interscience*, 1998.

[74] X. H. Wang, B. Zheng, W. F. Good, J. L. King, and Y. H. Chang. Computer-assisted diagnosis of breast cancer using a data-driven Bayesian belief network. *International Journal of Medical Informatics*, 54(2):115–126, 1999.

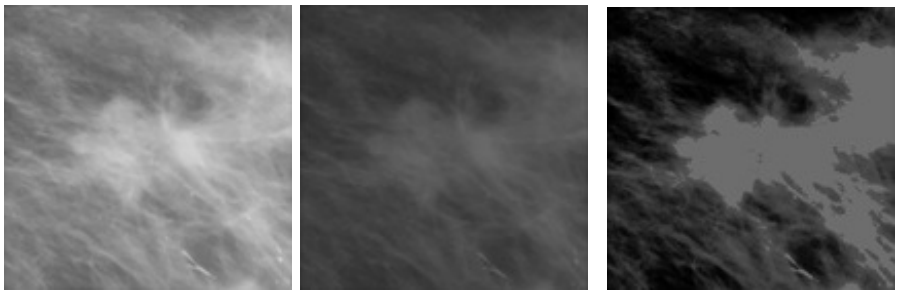
[75] Dong-Hui Xu, Arati S Kurani, Jacob D Furst, and Daniela S Raicu. Run-length encoding for volumetric texture. *Heart*, 27:1–25, 2004.

Anexo A

Técnicas de preprocesamiento

A.1. Preprocesamiento de imágenes: Ecualización y cuantización

El proceso de ecualización y cuantización en imágenes es importante para resaltar regiones de interés y permite que los descriptores que se obtengan a partir de estas imágenes proporcionen mayor información [30]. En este proyecto de tesis se utilizaron estas técnicas para mejorar el proceso de la extracción de las características de intensidad y textura. En la Figura A.1 se muestra un ejemplo del proceso de ecualización y cuantización.



(a) (b) (c)

Figura A.1: a) Imagen original, b) Imagen ecualizada y c) Imagen cuantizada.

Ecualización: Es una técnica de transformación no lineal que opera sobre los píxeles de una imagen de entrada y busca producir una imagen de salida con un histograma uniforme. La ecualización es un procedimiento que distribuye los niveles de gris del histograma de una imagen, utilizando de la mejor manera posible, el rango disponible de niveles de gris en la imagen de salida [30, 55]. Un ejemplo de una imagen ecualizada se puede ver en la Figura A.1b.

Cuantización: La cuantización de la imagen asignará a cada localización discreta (x, y) un valor entero 2^b, con b siendo: 1, 2, 3, 4, 5, 6, 7 o 8 bits por píxel [30]. Para la extracción de características de textura e intensidad que se propone en este proyecto de tesis se cuantizaron las imágenes de mamografía con un valor de $b = 7$. En la Figura A.1c se muestra una imagen cuantizada.

A.2. Preprocesamiento de datos

A.2.1. Normalización

En muchas situaciones prácticas se presentan datos cuyos valores se encuentran dentro de rangos dinámicos diferentes. Por lo tanto, las características con valores grandes pueden tener una mayor influencia en la función de costo de características con valores pequeños. Esto se puede tratar mediante la normalización de las características, de tal manera que los valores se encuentren dentro de rangos similares. Una técnica sencilla es la normalización a través de las respectivas estimaciones de la media y la varianza [70]. Para N datos de la $k - ésima$ característica la normalización se obtiene de la siguiente manera:

$$\hat{x}_{ik} = \frac{x_{ik} - \bar{x}_k}{\sigma_k} \tag{A.1}$$

Donde la media se obtiene como sigue:

$$\bar{x}_k = \frac{1}{N} \sum_{i=1}^{N} x_{ik}, \qquad k = 1, 2, ..., l \tag{A.2}$$

Y la varianza se obtiene de la siguiente manera:

$$\sigma_k^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_{ik} - \bar{x}_k)^2 \tag{A.3}$$

A.2.2. Discretización

Las redes Bayesianas pueden operar con valores continuos o discretos, sin embargo, normalmente operan con valores discretos. Los datos continuos pueden ser enteros o fracciones y los datos discretos son números enteros.

Los métodos de discretización se dividen en dos tipos principales, los no supervisados y los supervisados. Los métodos no supervisados no consideran la variable clase, así que las características con valores continuos son discretizados independientemente. El método no supervisado más simple es Discretización de Igual Anchura. Los métodos supervisados consideran la variable clase, es decir los puntos de división para formar rangos en cada atributo son seleccionados en función del valor de la clase [62].

El método de Discretización de Igual Anchura (*Equal Width Discretization EWD*), divide el rango de valores por cada atributo, $[X_{min}; X_{max}]$ en k intervalos, donde k es dado por el usuario u obtenido usando una cierta medida de información sobre los valores de los atributos. Este método clasifica los valores de v de cada característica en orden ascendente desde v_{min} a v_{max} en intervalos de tamaño igual a k. Cada intervalo tiene una anchura $w = (v_{min} - v_{max})/k$ y los puntos de corte se encuentran en $v_{min} + w, v_{min} + 2w, ..., v_{min} + (k-1)w$ [62].

Anexo B

Algoritmos para el aprendizaje de redes Bayesianas

B.1. Naïve Bayes aumentado a árbol (*Tree Augmented Naïve Bayes - TAN*)

En este trabajo se utilizó el algoritmo propuesto por Friedman *et al.*[26] para encontrar la topología de redes Bayesianas TAN. Este algoritmo contempla los siguientes pasos:

1. Calcular $I_n(X_i, X_j|C)$ para cada par de variables predictoras, con $i \neq j$.

2. Construir un grafo no dirigido completo en el cual los vértices son las variables predictoras $X_1, ..., X_n$. Asignar a cada arista conectando las variables X_i y X_j un peso dado por $I_n(X_i, X_j|C)$.

3. Construir un árbol expandido de máximo peso.

4. Transformar el árbol resultante no dirigido en un dirigido, escogiendo una variable raíz, y direccionando todas las aristas partiendo del nodo raíz.

5. Construir un modelo TAN añadiendo un nodo etiquetado como C, posteriormente un arco desde C a cada variable predictora X_i.

B.2. Clasificador Bayesiano K-Dependiente (*K-Dependence Bayesian Classifier - KDB*)

El método propuesto por Sahami [60] fue aplicado para encontrar la topología de redes Bayesianas KDB. Los pasos de este algoritmo son:

1. Para cada atributo X_i, se calcula la información mutua $I(X_i; C)$, con la variable clase.

2. Se calcula la información mutua condicionada a la clase $I(X_i; X_j|C)$ para cada par de atributos X_i y X_j, donde $i \neq j$.

3. Inicializar vacía a la lista S de variables usadas.

4. Se empieza a construir la red Bayesiana BN con un sólo nodo, la clase C

5. **mientras** S no incluya todos los atributos **repetir**

   a) Seleccionar el atributo X_{max} que no esté en S y que tenga el mayor valor de $I(X_{max}, C)$

   b) Añadir un nodo a la BN representando a X_{max}

   c) Añadir un arco desde C a X_{max} en la BN

   d) Añadir $m = min(|S|, k)$ arcos de m atributos distintos X_j en S hasta X_{max} con el valor más alto de $I(X_{max;X_j|C)}$

   e) Añadir X_{max} a S

6. **fin mientras**

7. Calcular las tablas de probabilidades condicionadas inferidas por la estructura del BN a partir de la lista de casos.

B.3. Naïve Bayes aumentado a bosque (*Forest Augmented Naïve Bayes - FAN*)

Para encontrar la topología de redes Bayesianas FAN, se utilizó el algoritmo propuesto por Lucas [44], el cual contempla los siguientes pasos:

1. Calcular la información mutua condicional $I_p(A_i; A_j|C)$, $j \neq i$ entre cada par de atributos, y calcular el promedio de la información mutua condicional I_{avg}.

2. Construir un grafo completo no dirigido, donde los nodos son atributos A_i, $i = 1, 2, ..., n$. Anotar el peso de una arista que hace conexión de A_j a A_i por $I_p(A_i; A_j|C)$.

3. Construir un árbol de expansión de peso máximo.

4. Calcular la información mutua $I_p(A_i; C)$, $i = 1, 2, ..., n$ entre cada atributo y la clase, y encontrar el atributo A_{root} que tiene la mayor información mutua con la clase.

5. Transformar el grafo dirigido resultante a uno dirigido, estableciendo A_{root} como la raíz y establecer las direcciones de todos los arcos hacia afuera de ella.

6. Eliminar los arcos dirigidos con el peso de la información mutua condicional menor que el promedio de la información mutua condicional I_{avg}.

7. Construir un modelo FAN añadiendo un arco etiquetado por C y la dirección de un arco dirigido desde C a cada A_i, $i = 1, 2, ..., n$.

Anexo C

Manual de usuario del programa desarrollado

El software de este trabajo de tesis se implementó en el lenguaje de programación Java (IDE Eclipse Ver. 4.2.1). La red Bayesiana codificada en el programa fue el modelo TAN que se describe en la Sección 5.5. Los detalles de instalación y utilización se presentan en las siguientes secciones.

C.1. Proceso de instalación

El software del clasificador de masas se distribuye como un archivo en formato **jar** –**clasificadorMasa.jar**. Para la ejecución de este archivo, es necesario tener instalado la Máquina Virtual de Java (MVJ) que se puede descargar del sitio: *https://www.java.com/es/download/*. El programa podrá ejecutarse al hacer doble clic sobre el archivo **clasificadorMasa.jar**.

C.2. Utilización del Software

El programa requiere para su ejecución de las siguientes imágenes en formato **pgm**:

- Masa: Es una imagen que contiene la masa a clasificar. El centro de la imagen debe corresponder al centro de la lesión (ver Figura C.1a). A esta imagen se le puede asignar cualquier nombre, por ejemplo:

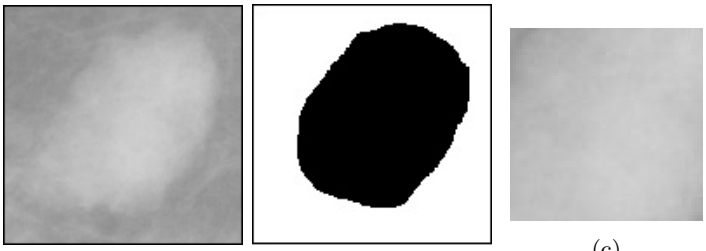
  nombreImagen.pgm

- Máscara de segmentación: Es una imagen que indica cuáles son los píxeles que corresponden a la masa. Los valores de los píxeles de la masa deben tener una intensidad de 0 y los del fondo 255 (Figura C.1b). Esta imagen debe tener las mismas dimensiones que la de masa y el nombre también debe ser el mismo más el prefijo "_msk", por ejemplo:

  nombreImagen_msk.pgm

- Región central: Es una imagen que contiene una muestra de la parte central de la masa. Su tamaño debe ser más pequeño que la imagen de masa, se recomienda que sea de 64×64 píxeles (Figura C.1c). El nombre de esta imagen debe ser el mismo a la de masa más el prefijo "_dsd", por ejemplo:

  nombreImagen_dsd.pgm



(a) (b) (c)

Figura C.1: Imágenes de a) la masa, b) máscara de segmentación y c) región central que requiere el programa de clasificación de masas.

C.2.1. Interfaz del programa

En la Figura C.2 se muestra la interfaz principal de este programa. La funcionalidad de cada uno de sus botones se describe a continuación:

- **Abrir imagen**: Este botón despliega la pantalla de la Figura C.3, permite explorar el directorio de la ubicación de la imagen de la masa y seleccionarla.
- **Extraer características**: Este botón se habilita después de seleccionar la imagen de la masa y permite calcular las características que se muestran en la sección de **Datos** de la Figura C.2.

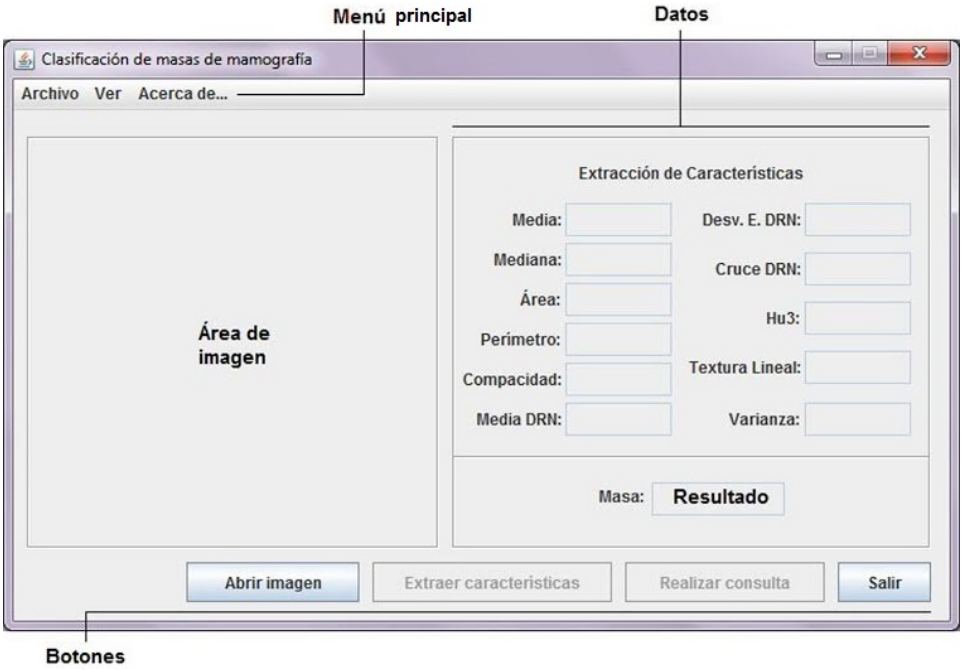


Figura C.2: Interfaz principal del programa.

- **Realizar consulta**: Este botón se habilita después de obtener las características, y proporciona la clasificación de la masa benigna o maligna.
- **Salir**: Cierra el programa de clasificación de masas de mamografía.

Las opciones del **Menú principal** que se muestran en la Figura C.2 tienen la siguiente funcionalidad:

- **Archivo**: Despliega los submenús **Abrir imagen** y **Salir**.

  Abrir imagen: Tiene la misma funcionalidad del botón **Abrir imagen**.

  Salir: Cierra el programa de clasificación de masas de mamografía.

- **Ver**: Despliega los submenús **Matriz topología**, **Topología** y **Tabla probabilidad**.

Matriz topología: Muestra la matriz de adyacencia de la red Bayesiana TAN implementada en este programa de clasificación (Figura C.4).

Topología: Muestra la topología de la red Bayesiana TAN (Figura C.5).

Tabla probabilidades: Muestra las tablas de probabilidades de la red Bayesiana TAN (Figura C.6).

- **Acerca de...**: Despliega el submenú **Versión**.

  Versión: Muestra la versión del programa (Figura C.7).

El botón **Cerrar** de las figuras C.4, C.5 y C.6, cierra la pantalla actual, dejando abierta la interfaz principal.

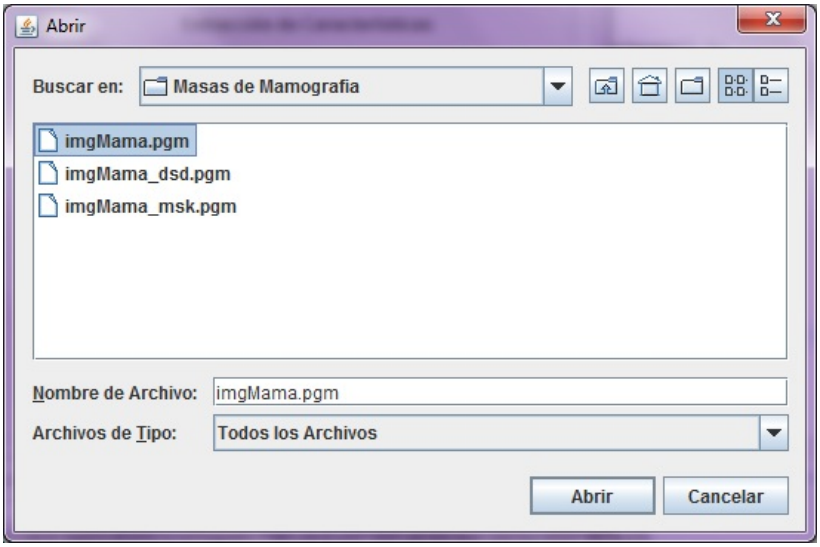
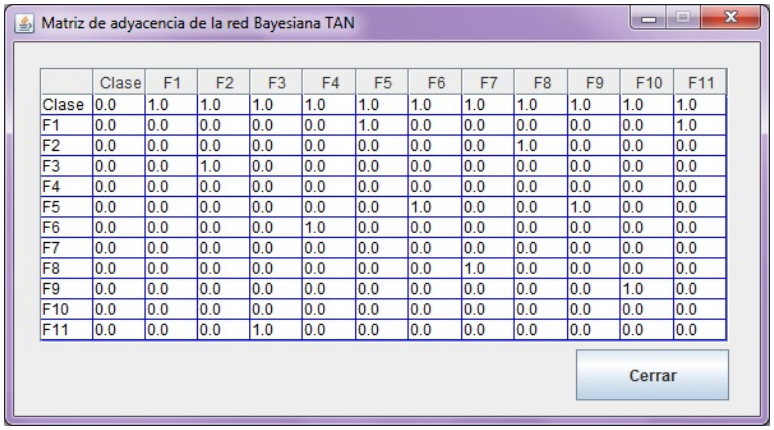


Figura C.3: Explorador de archivos para seleccionar la imagen de la masa.



	Clase	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11
Clase	0.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
F1	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0
F2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
F3	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
F4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
F5	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0
F6	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
F7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
F8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
F9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
F10	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
F11	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Figura C.4: Matriz de adyacencia de la red Bayesiana TAN.

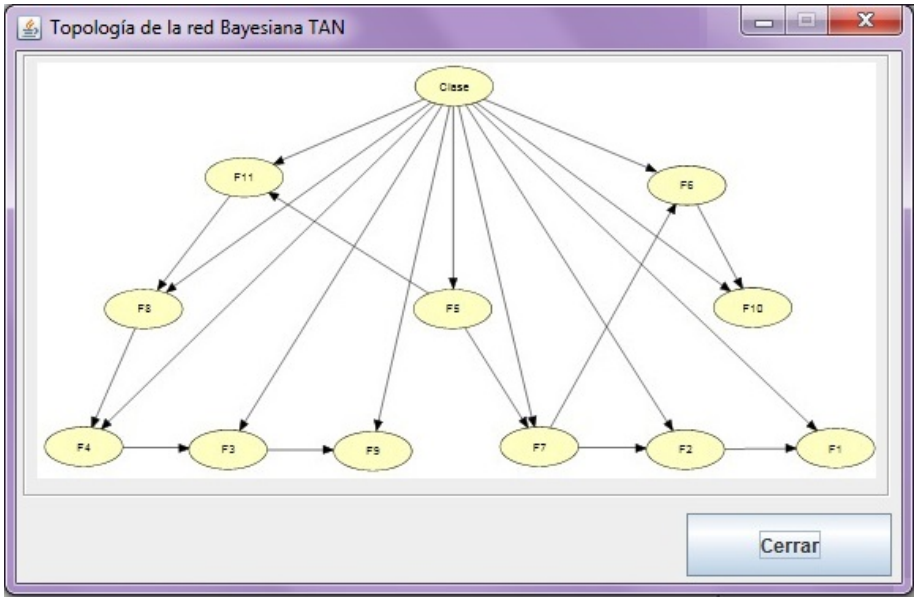


Figura C.5: Topología de la red Bayesiana TAN.

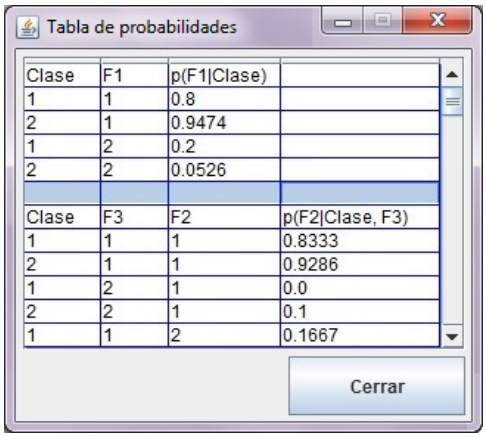


Figura C.6: Tablas de probabilidades.

Figura C.7: Versión del programa.

C.2.2. Clasificación de masa

El proceso a seguir para obtener la clasificación de una masa, son los siguientes:

a) Se selecciona la imagen de la masa, presionando el botón **Abrir imagen** (o la opción **Abrir imagen** del **Menú principal**). Con esta acción se desplegará la pantalla que se muestra en la Figura C.3, en el cual se podrá localizar la imagen de la masa a clasificar –imgMama.pgm, por ejemplo–. La imagen seleccionada se colocará en el recuadro **Área de imagen** y automáticamente se cargarán las imágenes correspondientes de la máscara de segmentación y de la región central.

b) El siguiente paso es extraer las características, presionando el botón **Extraer características**. Con esta acción se calcularán los descriptores que se muestran en la sección de **Datos**.

c) Finalmente, se determina la clasificación de la masa presionando el botón **Realizar consulta**. El resultado se desplegará en la sección de **Resultados** (Figura C.8).

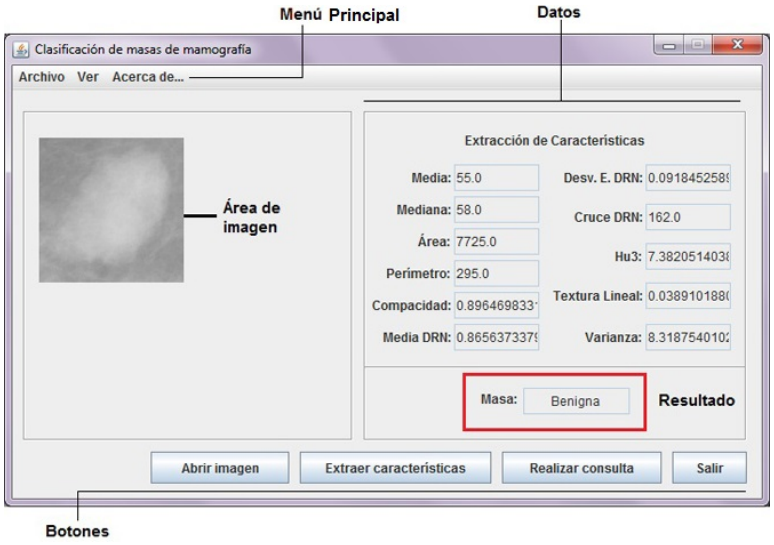


Figura C.8: Programa de clasificación de masas de mamografía.