INSTITUTO TECNOLÓGICO Y DE ESTUDIOS SUPERIORES DE MONTERREY
CAMPUS MONTERREY

ESCUELA DE INGENIERÍA Y TECNOLÓGIAS DE INFORMACIÓN

PROGRAMA DE GRADUADOS

Maestría en Ciencias con especialidad en Sistemas Inteligentes

Selección de características y clasificación de masas por medio de redes neuronales, máquinas de vector de soporte, análisis discriminante y regresión logística en mamografías digitales

Por

José Luis Hernández Cruz

ABRIL 2014

Selección de características y clasificación de masas por medio de redes neuronales, máquinas de vector de soporte, análisis discriminante y regresión logística en mamografías digitales

TESIS

Maestría en Ciencias con especialidad en Sistemas Inteligentes

Instituto Tecnológico y de Estudios Superiores de Monterrey

Por

José Luis Hernández Cruz

Abril 2014

Instituto Tecnológico y de Estudios Superiores de Monterrey

Campus Monterrey

Escuela de Ingeniería y Tecnologías de Información

Programa de Graduados

Los miembros del comité de tesis recomendamos que la presente tesis de José Luis Hernández Cruz sea aceptada como requisito parcial para obtener el grado académico de **Maestro en Ciencias** con especialidad en **Sistemas Inteligentes**.

Comité de Tesis:

Dr. Hugo Terashima Marín
Asesor Principal

Dr. Santiago E. Conant Pablos
Sinodal

Dr. José Gerardo Tamez Peña
Sinodal

Dr. Ramón F. Brena Piñero
Director de la Maestría en Ciencias con especialidad en Sistemas Inteligentes

Abril 2014

Selección de características y clasificación de masas por medio de redes neuronales, máquinas de vector de soporte, análisis discriminante y regresión logística en mamografías digitales

Por

José Luis Hernández Cruz

Presentada al Programa de Graduados de la Escuela de Ingeniería y Tecnologías de Información como requisito parcial para obtener el grado académico de

Maestro en Ciencias con especialidad en
Sistemas Inteligentes

Comité de Tesis:

Dr. Hugo Terashima Marín
Dr. Santiago E. Conant Pablos
Dr. José Gerardo Tamez Peña

Instituto Tecnológico y de Estudios Superiores de Monterrey
Campus Monterrey

Abril 2014

Agradecimientos

Quiero agradecer al café por las innumerables horas de compañía durante la realización del presente trabajo.

JOSÉ LUIS HERNÁNDEZ CRUZ

Instituto Tecnológico y de Estudios Superiores de Monterrey
Abril 2014

Thanks for all your unconditional confidence, support, patience, and encouragement.
You were my main motivation for pushing through this work.

Selección de características y clasificación de masas por medio de redes neuronales, máquinas de vector de soporte, análisis discriminante y regresión logística en mamografías digitales

by

José Luis Hernández Cruz

Abstract

El presente documento reporta el trabajo de tesis realizado por el autor, como miembro del Programa de Posgrados de la Escuela de Ingeniería y Tecnologías de Información, para la obtención del grado de Maestro en Ciencias con especialidad en Sistemas Inteligentes, bajo la supervisión del Dr. Hugo Terashima Marín.

El cáncer de mama es uno de los tipos de cáncer más comunes en la actualidad y posee tasas altas de mortalidad entre las mujeres a nivel mundial. El cáncer inicia cuando células anormales crecen sin control. En algunos casos, las células cancerosas forman un tumor y, si la enfermedad progresa, pueden trasladarse a otras partes del cuerpo generando metástasis.

La mamografía es una técnica no invasiva que permite identificar lesiones contenidas en el seno y que habilita así el diagnóstico de cáncer de mama. Una masa es un bulto o protuberancia, que puede ser palpable, que podría indicar la presencia de cáncer. La mamografía es uno de los exámenes recomendados para la detección temprana de masas en seno.

La presente investigación abarca la clasificación automática de masas en mamografías digitales, y propone un modelo de solución basado en fases de preprocesamiento, segmentación, extracción y selección de características y clasificación.

Inicialmente se detalla el proceso de construcción de un conjunto de ejemplos de masas tumorales a partir de las mamografías digitales de la base de datos *Digital Database for Screening Mammography (DDSM)*. Se extrae la información proporcionada por la base datos y se utiliza para la *segmentación* de masas a las cuales se le *extrae* un conjunto de 50 características. Estos ejemplos son usados para la construcción y validación del modelo propuesto.

La siguiente etapa consiste en la reducción de dimensiones mediante la selección de características. Primero se aplica el *análisis de correlación* y *ganancia de información* a las 50 características extraídas. Las características no excluidas en el análisis de correlación constituyen un espacio de búsqueda para el que se explora mediante un *algoritmo genético*. Los subconjuntos de características son evaluados con base en el rendimiento

del clasificador que considere únicamente a los elementos de ese subconjunto específico. Este modelo se conoce como de *envoltura* y en él se utilizan de forma separada cuatro algoritmos de clasificación: Redes neuronales (NN), Máquinas de vector de soporte (SVM), análisis discriminante (LDA) y Regresión logística (LR).

Una ve determinado el subconjunto de características que aportan mayor poder de discriminación, para algoritmo clasificador, se procede a la etapa de clasificación, donde se extraen estadísticas del desempeño de dichos algoritmos.

Al aplicar el modelo de solución propuesto sobre las imágenes de la base de datos DDSM, el algoritmo de máquinas de vector de soporte (SVM) obtuvo el mayor porcentaje de éxito con respecto a los otros tres algoritmos, tanto en la etapa de selección de características.

Índice de figuras

Índice de cuadros

Índice general

Capítulo 1

Introducción

Cáncer es un término genérico que designa un amplio grupo de enfermedades que pueden afectar a cualquier parte del organismo. Este grupo de enfermedades se caracteriza por la multiplicación de células anormales que se extienden más allá de sus límites habituales y que pueden invadir partes adyacentes del cuerpo o propagarse a otros órganos, proceso conocido como metástasis. Las metástasis son la principal causa de muerte en pacientes con cáncer.

El cáncer de mama es el tipo de cáncer más común entre las mujeres alrededor del mundo, siendo la principal causa de muerte por cáncer en mujeres a nivel global [1]. Acorde a la información de la Enciclopedia de Cáncer de Mama, 1 de 8 mujeres estadounidenses desarrollará cáncer de mama en algún momento de su vida. La misma fuente detalla que la incidencia entre las mujeres del Reino Unido es de 1 de entre cada 10 o 12 mujeres y en Australia es de 1 de cada 14 mujeres, mientras que en Japón la incidencia de cáncer de mama se duplicó entre 1960 y 1980 [58]. En México, entre 2004 y 2007, de cada 100 mujeres que fueron atendidas en instituciones de salud por padecer de algún tumor, 20 fueron diagnosticadas con cáncer de mama [32].

Las posibilidades que una mujer desarrolle cáncer de mama están relacionadas con diferentes factores de riesgo, algunos de los cuales pueden ser controlados como la lactancia materna, mantener una dieta saludable, ejercicio y cuidado del peso, evitar el consumo de alcohol y drogas, entre otros. Por otro lado, existen también factores como la historia familiar, actividad sexual, alteraciones genéticas y edad de la menarquia y menopausia que no están relacionados con el estilo de vida de las mujeres y por lo mismo no pueden ser alterados [41, 31].

No existe una forma completamente efectiva de prevención del cáncer de mama, pero un diagnóstico temprano y el consecuente tratamiento, aumenta las posibilidades de supervivencia de las personas diagnosticadas con esta enfermedad. Existen diferentes tipos de exámenes que permiten a los especialistas determinar la presencia de cáncer de mama en una mujer. De ellos, el más utilizado, principalmente por el potencial del mismo para realizarse periódicamente es la mamografía [53] .

La mamografía es una radiografía del seno que se puede usar para buscar el cáncer

de mama en mujeres que no presentan signos o síntomas de la enfermedad. Este procedimiento se elige según las características y preferencias de la mujer y del médico para buscar el cáncer de mama cuando hay o no síntomas. Por lo general, una mamografía requiere dos o más radiografías de cada seno. Las imágenes hacen posible que se detecten tumores que no se pueden palpar o encontrar lesiones, que conduzcan a la indicación de la presencia de cáncer de mama según el caso.

Las mamografías son interpretadas por radiólogos expertos, quienes emiten un diagnóstico con base en sus conocimientos y al contenido de las imágenes de cada paciente. Un método de diagnóstico que involucra el uso de computadoras para resaltar características de una mamografía y emitir un diagnóstico es conocido como Computer-Aided Diagnosis (CAD). Desde 1998 la U.S. Food and Drug Administration aprobó el primer dispositivo para CAD con base en imágenes médicas de seno [58]. Desde entonces se han incrementado los esfuerzos por desarrollar mejores y más precisas dispositivos.

El proceso de diagnóstico de cáncer asistido por computadora consiste de dos etapas. La primera de ella consiste en eliminar de la imagen aquellos elementos que no están relacionado con la posible y que podrían llegar a entorpecer el proceso de diagnóstico y resaltar aquellos elementos que podrían aportar al diagnóstico. La segunda etapa, es la etapa de clasificación de lesiones. En Aprendizaje de Máquina se dice que un problema es de clasificación cuando se tiene que decidir a que categoría pertenece una determinada entrada de un sistema. Todo sistema CAD cuenta con un clasificador, el cual es el responsable de decidir la lesión encontrada en la mamografía es maligna o benigna.

Para realizar la clasificación de lesiones se presenta un modelo que, como primer paso, toma el conjunto de características previamente extraídas de las mamografías y de estas, selecciona el subconjunto que aporte mayor poder de discriminación entre las lesiones. Una vez determinado este conjunto de características representativas, se sigue a la etapa de experimentación con los clasificadores seleccionados. Inicialmente se ha tomado en consideración dos clasificadores: Redes Neuronales (NN) [4], Máquinas de Vector de Soporte (SVM) [55] y Análisis de discriminante lineal (LDA) [25]. Con base en los experimentos se determinará el rendimiento de dichos clasificadores, así como el esquema de entrenamiento que aporte mejores resultados para cada algoritmos. El modelo termina con una etapa de validación y pruebas. Para que sea posible esta última etapa deberá realizarse una investigación previa sobre las diferentes bases de datos de mamografías, lo que permitirá comparar los resultados obtenidos en la investigación con diagnósticos previamente validados.

1.1. Definición del Problema

El cáncer es una enfermedad que se origina cuando las células en alguna parte del organismo comienzan a crecer de manera descontrolada. Estas células son conocidas como células cancerosas. El crecimiento de las células cancerosas es diferente al crecimiento de las normales. Cuando las células cancerosas llegan al torrente sanguíneo o a los vasos linfáticos, pueden propagarse hacia otras partes del cuerpo, en donde pueden continuar creciendo y formar nuevos tumores que invade el tejido normal. A este proceso se le llama metástasis y esta es la principal causa de muerte en pacientes que padecen cualquier tipo de cáncer.

Según datos de la Organización Mundial de la Salud [47], el cáncer es una de las principales causas de muerte en todo el mundo. Así, para el año 2008, 7.6 millones de defunciones fueron ocasionadas por algún tipo de cáncer y se prevé que las muertes por cáncer sigan una tendencia de crecimiento alrededor del mundo, estimando que para el año 2030 la cifra de decesos por cáncer alcance los 13.1 millones.

El cáncer es la segunda causa de muerte en los Estados Unidos [39]. En el año 2010, un estimado de 1529560 de personas fueron diagnosticadas con cáncer, de las cuales 569490 murieron como resultado del cáncer. Con base en las tendencias determinadas entre 2003 y 2005, más del 40 % de las personas que nacieron el día de hoy serán diagnosticadas con cáncer en algún momento de su vida. Los tipos de cáncer que más muertes causan cada año son los cánceres de pulmón, estómago, hígado, colon y mama.

Los tipos de cáncer más comunes entre hombres y mujeres son diferentes. El cáncer de mama es tipo de cáncer más común entre las mujeres alrededor del mundo, siendo la principal causa de muerte por cáncer en mujeres a nivel global. Acorde a la información de la Enciclopedia de Cáncer de Mama [58], 1 de 8 mujeres estadounidenses desarrollará cáncer de mama en algún momento de su vida. La misma fuente detalla que la incidencia entre las mujeres del Reino Unido es de 1 de entre cada 10 o 12 mujeres y en Australia es de 1 de cada 14 mujeres, mientras que en Japón la incidencia de cáncer de mama se duplicó entre 1960 y 1980. En México, entre 2004 y 2007, entre las mujeres que fueron atendidas en instituciones de salud por padecer de algún tumor, 20 de cada 100 fueron diagnosticadas con cáncer de mama [32].

GLOBOCAN muestra que el cáncer de mama tuvo, para el 2008, una incidencia total de 1384155, lo que equivale a 38.9 por cada 100000 mujeres, con una mortalidad de 458503, 12.4 por cada 100000 mujeres, siendo para ese año, el tipo de cáncer más común entre las mujeres alrededor del mundo [46].

En México la mortalidad por cáncer de mama ha mostrado un aumento notorio en las últimas cinco décadas. Entre 1955 y 1960, a partir de la disposición de los primeros datos confiables, la tasa era alrededor de dos a cuatro muertes por 100 000 mujeres. Luego se elevó de manera sostenida en las mujeres adultas de todas las edades hasta alcanzar una cifra cercana a 9 por 100000 para la mitad de la década de 1990 y se ha

mantenido estable desde entonces [33].

La mortalidad por cáncer de mama se puede reducir si los casos se detectan y tratan a tiempo. El diagnóstico temprano consiste en conocer los signos y síntomas iniciales para facilitar el tratamiento antes de que la enfermedad alcance una fase avanzada. Existen diferentes tecnologías que permiten realizar una detección temprana del cáncer de mama, entre estas sobresalen los exámenes clínicos de seno, mamografías, ultrasonidos y las resonancias magnética (MRI). De entre dichas técnicas, las mamografías continúan siendo la herramientas más utilizada para el diagnóstico, especialmente en países con economías en crecimiento, debido al bajo costo de las mismas y a la alta probabilidad que ofrecen al radiólogo para ofrecer un diagnóstico de cáncer.

La mamografía es una radiografía del seno. Las mamografías se pueden usar para buscar el cáncer de mama en mujeres que no presentan signos o síntomas de la enfermedad. Este tipo de mamografía se llama mamografía selectiva de detección; es decir, se elige este procedimiento según las características y preferencias de la mujer para buscar el cáncer de mama cuando no hay síntomas. Por lo general, una mamografía requiere dos radiografías o imágenes de cada seno. Las imágenes hacen posible que se detecten tumores que no se pueden palpar o encontrar lesiones, que conduzcan a la indicación de la presencia de cáncer de mama.

Las mamografías pueden usarse también para buscar el cáncer de mama después de haberse encontrado un abultamiento u otro signo o síntoma de dicho cáncer. Este tipo de mamografía se llama mamografía de diagnóstico. Los signos del cáncer de seno pueden ser dolor, engrosamiento de la piel, secreción del pezón o un cambio en el tamaño o forma del seno; sin embargo, estos signos pueden ser también signos de estados benignos. Una mamografía de diagnóstico puede usarse también para evaluar cambios que se encuentran durante una mamografía selectiva de detección o para ver el tejido del seno cuando es difícil obtener una mamografía de detección debido a circunstancias especiales como, por ejemplo, la presencia de implantes en los senos [43].

Las mamografías son interpretadas por radiólogos expertos, quienes son los encargados de realizar el diagnóstico con base en los aspectos presentes en las mamografías. Un radiólogo especialista puede detectar el cáncer de mama con base en radiografías con una tasa de falsos positivos igual a 10 % y de falsos negativos igual a 7 % [31]. A partir de lo anterior, es importante considerar que los humanos están son susceptibles a cometer errores, y que el análisis que realizan es subjetivo y cualitativo. Para aumentar la precisión en el diagnóstico, durante los último años, se ha echado mano de los sistemas computacionales, los cuales ayudan al radiólogo a realizar un diagnóstico objetivo y cuantitativo.

Un método de diagnóstico que involucra el uso de computadoras para resaltar características de una mamografía y emitir un diagnóstico es conocido como Computer-Aided Diagnosis (CAD). Desde 1998 la U.S. Food and Drug Administration aprobó el primer dispositivo para CAD con base en imágenes medicas de seno. Desde entonces se han incrementado los esfuerzos por desarrollar mejores y más precisas dispositivos.

Los principios de operación de las tecnologías CAD tienen como base las áreas de pre-procesamiento de la mamografía, extracción de características y clasificación de lesiones. En la etapa de pre-procesamiento de la mamografía, se eliminan de la imagen aquellos elementos que no tienen relación con el diagnóstico y que por el contrario podrían conducir errores en el mismo. La siguiente etapa consiste en extraer de la imagen un conjunto de características, las cuales se relacionan con diferentes aspectos encontrados en la imagen. Este conjunto de datos es la entrada de la siguiente etapa, detección y clasificación de lesiones. La etapa de clasificación es el objeto central de este estudio con base en cuatro aspectos fundamentales: reducción de características, algoritmos de clasificación, entrenamiento y pruebas.

El resultado de la etapa de extracción de características es un conjunto de elementos que identifican los aspectos encontrados en la mamografía. Para poder llevar a cabo el proceso de clasificación es necesario realizar una selección del total de características, reduciendo así en conjunto de entrada a un subconjunto que contiene los elementos del conjunto original que aportan la mayor precisión en el diagnóstico. Una vez determinado este conjunto de características a tener en cuenta se procede a la clasificación, para la cual el algoritmo de clasificación debe considerar lo aprendido con los casos similares, ya sea de manera previa o en la medida que se van presentando nuevos casos de mamografías. La presente propuesta de investigación tiene como objetivo desarrollar el proceso aquí descrito, validando y reportando al final del mismo los resultados obtenidos.

La presente propuesta de investigación esta principalmente motivada por el hecho que existe una necesidad por diseñar y desarrollar herramientas confiables para el diagnóstico de cáncer asistido por computadora. El desarrollo de estas herramientas aumentaría las posibilidades de supervivencia de las personas diagnosticadas con cáncer de mama, al ofrecer mayor precisión y confiabilidad en el diagnóstico de dicho padecimiento.

1.2. Motivación

El cáncer de mama es una enfermedad de gran magnitud y es considerada como uno de los principales problemas de salud pública en el mundo. La Organización Mundial de la Salud (OMS) considera que es una de las causas principales de muerte en la población. La posibilidad de curación y de mejora en la calidad de vida de las pacientes con cáncer de mama depende de la extensión de la enfermedad en el momento del diagnóstico [17].

Un diagnóstico a tiempo puede realizarse a través de la mamografía. El análisis de imágenes es un tema de importancia en el ámbito médico debido al incremento en el uso de equipos que obtienen imágenes médicas digitales directamente y, adicionalmente, por la accesibilidad a dispositivos para digitalización [18].

El encargado de analizar las mamografías o cualquier otro tipo de imagen médica es

el especialista en radiología. Este trabajo puede resultar complicado ya sea por la falta de experiencia, por trabajo acumulado y cansancio o por muchos otros factores humanos. El uso de herramientas denominadas *computer-aided diagnosis (CAD)* apoyan en la interpretación de las imágenes y brindan opciones para un diagnóstico más preciso [37].

El presente trabajo de tesis se realizó con la intención de apoyar el desarrollo de los sistemas CAD, específicamente en la su labor de clasificación de lesiones de masas tumorales. Durante el desarrollo de la tesis, se realizó una investigación y experimentación con algoritmos de aprendizaje automático en busca de alcanzar y mejorar el desempeño de los sistemas actuales.

1.3. Objetivos

El objetivo general de esta investigación es diseñar, probar y validar los clasificadores, Redes Neuronales (NN), Máquinas de Vector de Soporte (SVM), Análisis de Discriminante Lineal (LDA) y Regresión Logística (LR) para el proceso de diagnóstico de lesiones de masas tumorales en mama con base en características representativas previamente extraídas de mamografías digitales. Estos clasificadores serán capaces de decidir si una lesión es benigna o maligna, considerando para ello un conjunto de casos de aprendizaje previamente diagnósticados. Para alcanzar dicho objetivo se llevará a cabo un estudio de la teoría relacionada con el aprendizaje de automático y computación evolutiva, específicamente en cuanto a Redes Neuronales (NN), Máquinas de Vector de Soporte (SVM), Análisis de Discriminante Lineal (LDA) y Regresión Logística (LR).

Los objetivos particulares de esta investigación son:

- Construir una base de datos apropiada para el dominio de este problema y validación de resultados.
- Determinar el conjunto de características representativas que aporten mayor poder discriminativo para la correcta clasificación de masas en mamografías.
- Identificar y evaluar algoritmos para clasificar, entre benignas y malignas, las masas tumorales encontradas en mamografías digitales.
- Realizar una comparación entre los algoritmos seleccionados, Redes Neuronales (NN), Máquinas de Vector de Soport (SVM), Análisis de Discriminante Lineal (LDA) y Regresión Logística (RL) tiene el mejor desempeño para clasificar las masas detectadas en mamografías.

- Realizar una comparación entre los resultados obtenidos con los clasificadores Redes Neuronales (NN), Máquinas de Vector de Soport (SVM), Análisis de Discriminante Lineal (LDA) y Regresión Logística (RL) y los que se han usado en los trabajos relacionados.

1.4. Alcances y suposiciones

Los alcances de la presente investigación abarcan:

- Aplicar técnicas desarrolladas y probadas en trabajos previos para la fase de preprocesamiento y segmentación, extracción y selección de características y clasificación de lesiones de masas en mamografías digitales.
- Utilizar los los algoritmos de Redes Neuronales (NN), Máquinas de Vector de Soporte (SVM), Análisis de Discriminante Lineal (LDA) y Regresión Logística (RL) para la clasificación masas en malignas y benignas.
- Utilizar la base de datos *Digital Database of Screening Mammographies (DDSM)* para generar un conjunto de ejemplos que puedan ser utilizados para construir un modelo de clasificación de lesiones de masas tumorales.
- Probar efectividad del modelo propuesto, comparando los resultados entregados la etapa de clasificación mediante la matriz de confusión, precisión, sensibilidad, especificidad y puntaje F.

Las suposiciones sobre las que se apoya el presente trabajo son:

- La información sobre las anomalías en la base de datos DDSM es correcta.
- Las masas de las mamografías de la base DDSM se encuentran contenidas en el borde que se describe en el archivo asociado a cada hallazgo.

1.5. Hipótesis

Es posible determinar un algoritmo clasificador capaz de discriminar de manera automática entre masas tumorales benignas y malignas, a partir de la información extraída de de una base de datos de mamografías digitales previamente diagnosticadas.

Al mismo tiempo, es posible determinar mediante algoritmos evolutivos el conjunto de entradas de dicho clasificador que maximiza su rendimiento.

Las preguntas a responder con esta investigación son las siguiente:

- ¿Cuál es el conjunto de características que aporta el mayor poder discriminativo de las lesiones de masas en mamografías digitales?
- ¿Existe un algoritmo o una metodología para la clasificación de masas tumorales en mamografías que entrega resultados significativamente superiores al resto?
- ¿Cuál es el desempeño del algoritmo de clasificación en el diagnóstico asistido por computadora de masas tumorales que presenta mejores resultados?
- ¿Existen trabajos relacionados que aborden el problema de clasificación de masas en mamografías?
- ¿Qué parámetros influyen el desempeño de los diferentes algoritmos de clasificación de masas tumorales en mamografías?
- ¿Existen bases de datos con recursos que permitan validar la clasificación de lesiones de masas identificadas en mamografías?

1.6. Contribuciones

La presente investigación contribuye con un modelo de solución para la detección y clasificación de masas en mamografías digitales, aportando conocimientos sobre métodos y algoritmos para su ubicación y diagnóstico, de forma particular en los siguientes puntos:

- Construir, a partir de mamografías digitales de la *Digital Database of Screening Mammographies (DDSM)*, una base de datos de lesiones de masas tumorales.
- Determinar un conjunto de características que describan cuantitativamente una masa tumoral y que puedan ser utilizadas determinar de manera automática la malignidad de una lesión.
- Seleccionar experimentalmente, y mediante el algoritmos de búsqueda, un subconjunto de características que sirvan de entrada a un algoritmo clasificador.

- Realizar experimentación para analizar resultados obtenidos de la investigación y compararlos con trabajos previos.
- Comparar el desempeño de los cuatro algoritmos de clasificación para la clasificación de masas con base en las métricas de *éxito*, *sensibilidad*, *especificidad* y *error*.

1.7. Organización de la Tesis

La organización del presente documento de tesis, que describe la importancia de un método adecuado para la clasificación de masas en mamografías digitales, se detalla a continuación:

En el capítulo 2 se describen los antecedentes sobre cáncer de mama, las generalidades del área de aprendizaje automático, los detalles más importantes de los algoritmos de clasificación empleados en la investigación, las formas de evaluar del rendimiento de dichos algoritmos y algunos conceptos adicionales que permiten al lector una mejor comprensión del contenido de la tesis. El capítulo 3 detalla el modelo de solución planteado y describe cada una de sus fases y algoritmos desarrollados en la presente investigación. Los experimentos y sus análisis realizados constan en el capítulo 4, divididos en experimentos exploratorios para determinar valores de parámetros de cada algoritmo, y experimentos de validación que prueban el modelo de solución e hipótesis planteados. Para terminar, en el capítulo 5 se presentan las conclusiones obtenidas, se indican contribuciones y se listan algunas sugerencias para trabajos futuros.

Capítulo 2

Antecedentes

El presente capítulo del documento abarca los conceptos generales de las áreas relacionadas con este trabajo de investigación. Inicialmente se presenta una sección que provee la descripción del cáncer de mama como enfermedad, resaltando algunos factores de riesgo que contribuyen a su desarrollo y los síntomas que conlleva. El detalle de este padecimiento continua mediante la descripción de los diferentes tipos de cáncer de mama y las diferentes etapas en las que se divide su evolución. Posteriormente se listan los métodos comúnmente utilizados para detectar y diagnosticar este padecimiento, resaltando los aspectos relacionados con la mamografía, siendo esta la más utilizada y que contiene información que permite detectar el padecimiento incluso cuando la paciente no presenta síntomas.

El capítulo continua con la descripción de los aspectos computacionales considerados en el modelo de solución aquí descrito. El aprendizaje automático es un área de la inteligencia artificial donde los sistemas computacionales tienen como objetivo la generación de conocimiento mediante aprendizaje a partir de información y datos. Se incluye una descripción de los algoritmos de aprendizaje automático considerados en la presente investigación, resaltando las capacidades y limitaciones de los mismos y las estrategias para evaluar el rendimiento de algoritmos de este tipo. Posteriormente se incluye la información relacionada con las bases de datos que se utilizaron en el la validación del modelo. Al final del capítulo se muestran algunos trabajos de investigaciones afines a la presente.

2.1. Cáncer de Mama

El seno o mama está compuesto por lóbulos y conductos. Cada mama tiene entre 15 y 20 secciones que se llaman lobulillos. Los lobulillos terminan en docenas de bulbos de tamaño pequeño, los cuales pueden producir leche. Los lóbulos, lobulillos y bulbos están conectados por tubos delgados conocidos como conductos. Cada seno tiene, al mismo tiempo, vasos sanguíneos y vasos linfáticos. Los vasos linfáticos transportan un líquido

prácticamente incoloro al que se le denomina linfa [41]. Los vasos linfáticos conducen a órganos pequeños que se llaman ganglios linfáticos. Los ganglios linfáticos son estructuras pequeñas con forma de frijol que se encuentran en todo el cuerpo. Filtran sustancias de un líquido que se llama linfa y ayudan a combatir infecciones y enfermedades. Hay racimos de ganglios linfáticos cerca de la mama en las axilas (debajo de los brazos), por encima de la clavícula y en el pecho.

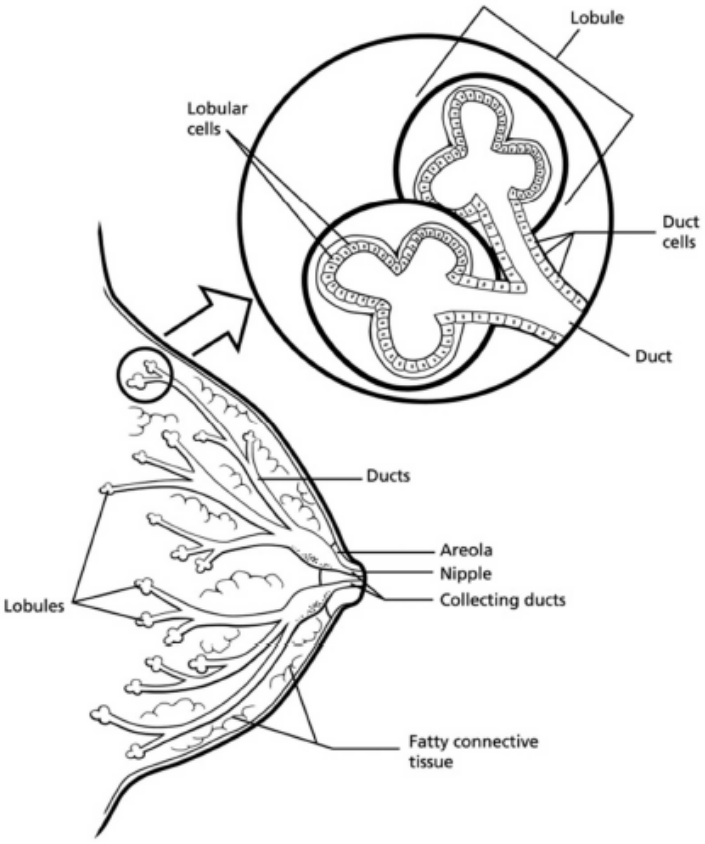


Figura 2.1: Anatomía de la mama femenina.

La figura 2.1, muestra, en la parte derecha, el pezón y la aréola en el exterior de la mama. También se muestran los ganglios linfáticos, los lóbulos, los lobulillos, los conductos y otras partes del interior de la mama. Por su parte la figura 2.2 muestra el sistema linfático del seno femenino.

El cáncer es una enfermedad que se origina cuando las células en alguna parte del organismo comienzan a crecer de manera descontrolada. Estas células son conocidas como células cancerosas. El crecimiento de las células cancerosas es diferente al crecimiento de las células normales. En lugar de morir, las células cancerosas continúan creciendo y formando más células cancerosas, las cuales pueden crecer hacia otros tejidos (invadir), algo que las células normales no hacen. La posibilidad de una célula de crecer sin control

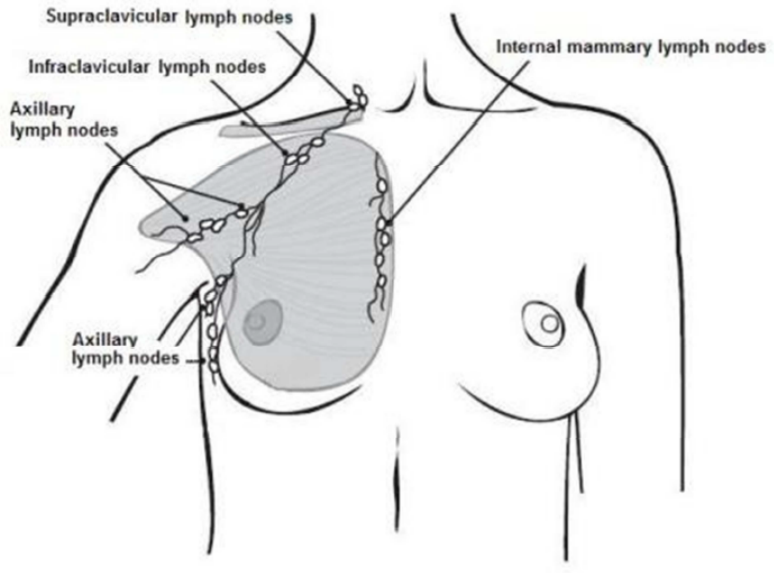


Figura 2.2: Sistema linfático del seno.

e invadir otro tejido es lo que la hace cancerosa.

Cuando las células cancerosas llegan al torrente sanguíneo o a los vasos linfáticos, pueden propagarse hacia otras partes del cuerpo, en donde pueden continuar creciendo y formar nuevos tumores que invaden el tejido normal. A este proceso se le llama metástasis y esta es la principal causa de muerte en pacientes que padecen cualquier tipo de cáncer.

El cáncer de mama consiste en el crecimiento descontrolado de células malignas en el seno. Esta enfermedad ocurre principalmente en las mujeres, pero los hombres también pueden padecer cáncer de seno [3].

2.1.1. Factores de Riesgo

Un factor de riesgo es cualquier situación o circunstancia que aumenta la probabilidad de desarrollar una enfermedad. Entre los factores de riesgo más comunes para presentar cáncer de mama, se cuentan los siguientes [43, 41]:

- **Edad:** El riesgo para una mujer de padecer cáncer de mama incrementa directamente con la edad de la mujer. 2 de cada 3 casos de cáncer de mama invasivo son detectados en pacientes con una edad mayor o igual a 55.
- **Historia personal:** Mujeres que han padecido cáncer de mama son muy propensas a desarrollarlo por segunda ocasión. Una mujer con cáncer en un seno tiene de 3 a 4 veces más riesgo de desarrollar cáncer en el mismo seno o en el otro.
- **Historia familiar:** La probabilidad que una mujer desarrolle cáncer de mama es mayor si algún familiar cercano, ya sea hombre o mujer, ha padecido la enfermedad, especialmente si esa persona fue diagnosticada antes de cumplir los 50

años. Una mujer que tiene un familiar de primer grado (madre, hermana o hija) que ha padecido cáncer de mama tiene un riesgo de padecer de la enfermedad de aproximadamente el doble de una mujer que no y el riesgo sube al triple si son dos familiares en primer grado quienes han padecido ésta enfermedad.

- **Alteraciones genéticas:** Cambios en algunos genes, como por ejemplo el BRCA1 y BRCA2, incrementan la probabilidad que una mujer padezca cáncer de mama.
- **Densidad mamaria:** Los senos o mamas están constituidos de tejido graso, fibroso y glandular.Se dice que una mujer tiene un tejido mamario denso (lo cual se observa en las radiografías de seno) cuando posee más tejido fibroso y glandular que graso. Mujeres con tejido mamario denso poseen mayor riesgo de padecer cáncer de mama.
- **Historia menstrual y reproductiva:** Mujeres que han tenido su primera menstruación antes de los 12 años o que han entrado a la menopausia después de los 55 años son altamente propensas a desarrollar cáncer de mama. Al mismo tiempo mujeres que han tenido su primer embarazo y han dado a luz después de los 30, o aquellas que han estado embarazadas y que no han llegado a dar a luz también tienen altas probabilidades de padecer cáncer de mama.
- **Lactancia:** Algunos estudios sugieren que la lactancia materna podría disminuir ligeramente el riesgo de padecer cáncer de mama, especialmente si la lactancia se mantiene de manera continua por periodos entre 1.5 y 2 años.
- **Consumo de alcohol:** Algunos estudios muestran que entre más alcohol consume una mujer, tiene mayor probabilidad de presentar cáncer de mama. Mujeres que consumen de 2 a 5 bebidas alcohólicas al día tienen 1.5 más riesgo de padecer cáncer de mama que mujeres que no consumen alcohol.
- **Peso:** Las probabilidades de sufrir cáncer de mama después de la menopausia son más altas en mujeres con sobrepeso u obesas.
- **Actividad física:** Mujeres con un estilo de vida con poca actividad física pueden tener mayor riesgo de cáncer de mama. La actividad física y el ejercicio ayudan a reducir las posibilidades de desarrollar la enfermedad. En un estudio de la Iniciativa para la Salud de la Mujer (Women's Health Initiative) muestra que de 1.25 a 2.5 horas de caminata ligera reduce el riesgo de cáncer de mama en un 18 %.
- **Raza y etnicidad:** La incidencia del cáncer de seno es más alta en las mujeres blancas para la mayoría de los grupos de edad. Sin embargo, las afroamericanas presentan índices de incidencia más elevados antes de los 40 años de edad e índices de mortalidad más altos que cualquier otro grupo racial o étnico de los Estados Unidos a cualquier edad. La brecha en mortalidad entre las mujeres afroamericanas

y las blancas es ahora más grande que a principios de los años noventa. Las mujeres con origen asiático, hispano o nativas americanas tienen un riesgo menor de padecer y morir por causa del cáncer de mamá.

- **Control de natalidad:** Estudios han encontrado que mujeres que utilizan anticonceptivos de vía oral, como píldoras, tienen un riesgo ligeramente mayor de padecer cáncer de mama que aquellas mujeres que nunca han utilizado dicho de método.
- **Exposición a radiación**: Mujeres que han recibido terapia de radiación en área del pecho como tratamiento para algún tipo de cáncer tienen un riesgo mayor de padecer cáncer de mama.

2.1.2. Síntomas

Acorde con el National Cancer Institute de los Estados Unidos, el cáncer de mama puede presentarse por medio de cualquiera de los siguientes signos y síntomas [43]:

- Masa o engrosamiento en la mama o cerca de ella, o en el área debajo del brazo.
- Cambio en el tamaño o la forma de la mama.
- Hueco o arruga en la piel de la mama.
- Pezón que se vuelve hacia adentro de la mama.
- Líquido que sale del pezón, que no es leche materna; especialmente si es sanguinolento.
- Piel con escamas, roja o hinchada en la mama, el pezón o la areola (área oscura de piel que rodea el pezón).
- Huecos en la mama parecidos a la piel de la naranja.

Estos síntomas son los principalmente asociados con el cáncer de mama, sin embargo no son exclusivos de esta enfermedad, otro tipo de patologías también pueden presentar los mismos síntomas.

2.1.3. Tipos de Cáncer de Mama

Carcinoma es el término que se usa para describir un cáncer que ha comenzado en la capa de revestimiento de órganos como la mama. *Carcinoma in situ* significa que las células cancerosas están solamente en los conductos o en los lobulillos, no se ha propagado al tejido más profundo en la mama ni a otros órganos del cuerpo. *Carcinoma invasivo* es un tipo de cáncer donde la enfermedad ha alcanzado más allá de la capa

de células donde se originó. Existen muchos tipos de cáncer de seno, pero algunos de ellos ocurren muy pocas veces. Los principales tipos de cáncer de mama se describen a continuación [3, 43]:

1. **Carcinoma ductal in situ:** éste es el tipo más común de cáncer no invasivo de mama. El carcinoma ductal in situ (ductal carcinoma in situ, DCIS) significa que el cáncer está solamente en los conductos y no se ha propagado a través de las paredes de los conductos al tejido del seno. Por lo tanto, no se puede propagar a los ganglios linfáticos u a otros órganos.

2. **Carcinoma lobulillar in situ:** en realidad, el carcinoma lobulillar in situ no es un tipo cáncer, pero su presencia indica un riesgo mayor en la aparición de cáncer de mama.

3. **Carcinoma ductal invasivo (o infiltrante):** éste es el cáncer de mama más común. El carcinoma ductal invasivo (invasive ductal carcinoma, IDC) comienza en un canal (conducto) lácteo, penetra la pared del conducto e invade el tejido del seno. Desde ese lugar es posible que pueda propagarse a otras partes del cuerpo (hacer metástasis). Es responsable de aproximadamente ocho de cada 10 casos de cáncer invasivo del seno.

4. **Carcinoma lobulillar invasivo (infiltrante):** este cáncer se origina en las glándulas mamarias (lobulillos) y luego se propaga a través de la pared de los lobulillos. El carcinoma lobulillar invasivo (invasive lobular carcinoma, ILC) puede propagarse a otras partes del cuerpo. Aproximadamente 1 de 10 casos de cáncer invasivo del seno son de este tipo.

5. **Cáncer inflamatorio del seno:** este tipo de cáncer invasivo del seno no es común. Representa aproximadamente de 1 a 3 % de la totalidad de casos de cáncer de mama. Por lo general, no se presenta una sola protuberancia o tumor, sino que este cáncer hace que la piel del seno luzca rojiza y se sienta acalorada. El cáncer inflamatorio del seno (inflammatory breast cancer, IBC) también puede hacer que la piel se haga más gruesa y presente hoyuelos, como la cáscara de una naranja. Puede que el seno se vuelva más grande, duro, sensible o que sienta picazón.

Otros tipos de cáncer de mama poco comunes son: Cáncer de seno triple negativo, tumores mixtos, carcinoma medular, carcinoma metaplásico, carcinoma mucinoso, enfermedad de Paget del pezón, carcinoma tubular, carcinoma papilar, carcinoma quístico adenoide (carcinoma adenoquístico), tumor loides, y angiosarcoma.

2.1.4. Etapas de Cáncer de Mama

La estadificación describe la gravedad del cáncer que aqueja a una persona basándose en la extensión del tumor original (primario) y si el cáncer se ha diseminado en el

cuerpo o no. La estadificación se basa en los conocimientos sobre el proceso de evolución del cáncer. En [43] se resaltan los siguientes aspectos estadificación del cáncer:

- La estadificación ayuda al médico a planear un tratamiento apropiado.
- La etapa o estadio puede usarse para estimar el pronóstico de la persona.
- Conocer la etapa es importante para identificar estudios clínicos que puedan ser adecuados para un paciente en particular.
- La estadificación permite a los profesionales médicos e investigadores compartir información sobre los pacientes.
- Facilita un lenguaje común para evaluar los resultados de los estudios clínicos y comparar resultados de estudios diferentes.

El sistema TNM es uno de los sistemas de estadificación de mayor uso. Este sistema ha sido aceptado por la International Union Against Cancer, UICC, y por el American Joint Committee on Cancer, AJCC. La mayoría de los establecimientos médicos usan el sistema TNM como método principal de reportar sobre el cáncer.

El sistema TNM está basado en la extensión del tumor (T, ver tabla 2.1), el grado de diseminación a los ganglios linfáticos (N, ver tabla 2.2), y la presencia de metástasis (M, ver tabla 2.3) distante. Un número se añade a cada letra para indicar el tamaño o extensión del tumor y el grado de diseminación del cáncer.

Después de que se diagnostica el cáncer de mama, se realizan pruebas para determinar si las células cancerosas se diseminaron dentro de la mama o hasta otras partes del cuerpo. Para el cáncer de mama se usan los siguientes estadios [43]:

1. **Estadio 0:** Este es normalmente conocido como carcinoma no invasivo o carcinoma in situ.
2. **Estadio I:** Esta es una etapa temprana del cáncer de mama en donde se ha expandido a través de los ductos o lóbulos e invadido el tejido mamario circundante. Estadio I significa que el tamaño del tumor es menor a 1 centímetro y que el cáncer no ha ido más allá del seno.
3. **Estadio II:** El cáncer ya se formá y se presenta alguno de los siguientes escenarios:
   - No se encuentra un tumor en la mama, pero se encuentra cáncer en los ganglios linfáticos axilares (ganglios linfáticos debajo del brazo)
   - El tumor mide dos centímetros o menos, y se diseminá hasta los ganglios linfáticos axilares
   - El tumor mide más de dos centímetros, pero no más de cinco centímetros, y no se disemine hasta los ganglios linfaticos axilares

	Tumor Primario
Tx	Tumor primario no puede ser evaluado
T0	No hay evidencia de tumor primario
Tis	CDIS Carcinoma Ductal in Situ CLIS Carcinoma Lobulillar in Situ Enfermedad de Paget del pezón
T1	Tumor $\leq$ 20mm
T1mi	Tumor $\leq$ 1mm en su diámetro mayor
T1a	Tumor $>$ 1mm pero $\leq$ 5mm en su diámetro mayor
T1b	Tumor $>$ 5mm pero $\leq$ 10mm en su diámetro mayor
T1c	Tumor $>$ 10mm pero $\leq$ 20mm en su diámetro mayor
T2	Tumor $>$ 20mm pero $\leq$ 50mm en su diámetro mayor
T3	Tumor $>$ 50mm en su diámetro mayor
T4	Tumor de cualquier tamaño con extensión directa a la pared torácica y/o dermis(ulceración o nódulos cutáneos). La invasión a la dermis, no se considera como T4
T4a	Extensión a la pared torácica, no incluye solo la adherencia o invasión al músculo pectoral
T4b	Ulceración y/o nódulos satélite y/o edema (incluye piel de naranja) de la piel, que no cumple con criterios de carcinoma inflamatorio
T4c	T4a y T4b combinados
T4d	Carcinoma inflamatorio

Cuadro 2.1: Sistema TNM: Tumores

4. **Estadio III:** En esta etapa el tumor ha alcanzo un tamaño mayor a las anteriores y el cáncer se ha expandido a alguna de las siguiente partes del cuerpo:
   - Los nodos linfáticos debajo del brazo.
   - Los nodos linfáticos cerca del hueso del seno.
   - Otro tejido cercano a la mama.
5. **Estadio IV:** el cáncer se disemina hasta otros órganos del cuerpo, con mayor frecuencia hasta los huesos, los pulmones, el higado o el cerebro.

Los diferentes estadios según el sistema TNM se muestran en la tabla 2.4

2.1.5. Diagnóstico

El cáncer de mama en etapas iniciales se presenta de manera subclínica en la mayoría de los casos, es decir que solamente es detectable por estudios de imagen (mastografía, ultrasonido y resonancia magnética), en menor proporción por clínica (tumores palpables); sin embargo otra forma de presentación común es como un tumor no doloroso que hasta en 30 % se asocia con trastornos en los ganglios linfáticos de la axila [7].

	Ganglios
Nx	Los ganglios regionales no pueden ser evaluados
N0	No hay metástasis en los ganglios regionales
N1	Metástasis móviles en los ganglios ipsilaterales en el nivel I o II de la axila
N2	Metástasis en los ganglios ipsilaterales en el nivel I o II de la axila, que están fijos o en el conglomerado. Ganglios positivos en la cadena mamaria interna, en ausencia de ganglios axilares clínicamente palpables
N2a	Metástasis en los ganglios de axilares del nivel I o II, fijos entre ellos o a otras estructuras
N2b	Metástasis en los ganglios de la cadena mamaria interna en ausencia de ganglios axilares clínicamente detectables
N3	Metástasis a ganglios infraclaviculares (nivel III) ipsilaterales con o sin involucrar ganglios de los niveles I o II. Ganglios en la cadena mamaria interna con afección de los ganglios del nivel I o II axilar. Metástasis en los ganglios supraclaviculares ipsilaterales con o sin afección en los ganglios axilares o de la cadena mamaria interna
N3a	Metástasis a ganglios infraclaviculares ipsilaterales
N3b	Metástasis a ganglios ipsilaterales de la cadena mamaria interna
N3c	Metástasis a ganglios supraclaviculares ipsilaterales

Cuadro 2.2: Sistema TNM: Ganglios

	Metástasis
M0	No hay evidencia clínica o radiográfica de metástasis a distancia
cM0 (i+)	No hay evidencia clínica o radiográfica de metástasis a distancia, pero existen depósitos moleculares o microscópicos detectados por células tumorales circulantes en sangre, médula ósea o ganglios reginales menores a 0.2mm en una persona sin síntomas de metástasis
M1	Metástasis a a distancia detectables

Cuadro 2.3: Sistema TNM: Metástasis

Para detectar y diagnosticar el cáncer de mama se utilizan comúnmente las siguientes pruebas[43]:

- **Examen físico y antecedentes:** examen del cuerpo para revisar los signos generales de salud, incluso verificar si hay signos de enfermedad, como masas o cualquier otra cosa que parezca anormal. También se anotan los antecedentes de los hébitos de salud del paciente y los antecedentes médicos de sus enfermedades y tratamientos anteriores.
- **Mamografía:** La mamografía, a veces denominanda mastografía, es hasta ahora el mejor método de detección, tiene una sensibilidad diagnóstica de 80 a 95 %,

Estadios	T	N	M
0	Tix	N0	M0
I	T1	N0	M0
IIA	T0	N1	M0
	T1	N1	M0
	T2	N0	M0
IIB	T2	N1	M0
	T3	N0	M0
IIIA	T0	N2	M0
	T1	N2	M0
	T2	N2	M0
	T3	N1	M0
	T3	N2	M0
IIIB	T4	N0	M0
	T4	N1	M0
	T4	N2	M0
IIIC	Cualquier T	N3	M0
IV	Cualquier T	Cualquier N	M1

Cuadro 2.4: Agrupación TNM

aunque 10 a 15 % de los tumores puede ser oculto principalmente en mujeres con mamas densas (con el uso de mamografía digital mejora la sensibilidad diagnóstica en este grupo de pacientes).

- **Ultrasonido:** Conocido también como ecografía, es un procedimiento en el que se hacen rebotar ondas de sonido de alta energía (ultrasonidos) en los tejidos u órganos internos para producir ecos. Los ecos se utilizan para construir una imagen de los tejidos corporales. El ultrasonido es en algunos casos una herramienta complementaria para diferenciar masas quísticas de sólidas, para caracterizar lesiones benignas y malignas y como guía para la realización de biopsias de lesiones no palpables.

- **IRM (imágenes por resonancia magnética):** procedimiento para el que se usa un imán, ondas de radio y una computadora para crear imágenes detalladas de áreas internas del cuerpo. Este procedimiento también se llama imágenes por resonancia magnética nuclear (IRMN). La imagen por resonancia magnética (IRM) con gadolinio tiene sensibilidad diagnóstica de 94 a 100 %, pero baja especificidad (37 a 97 %) y valor predictivo positivo de 44 a 96 %.

- **Estudios químicos de la sangre:** procedimiento por el cual se examina una muestra de sangre para medir las cantidades de ciertas sustancias que los órganos y tejidos del cuerpo liberan en la sangre. Una cantidad anormal (mayor o menor que la normal) de una sustancia puede ser signo de enfermedad en el órgano o el tejido que la elabora.

- **Biopsia:** extracción de células o tejidos para que un patólogo pueda observarlas bajo un microscopio y verificar si hay signos de cáncer. Si se encuentra una masa en la mama, el médico puede necesitar extraer una pequeña cantidad de la masa.

2.1.6. Diagnostico Asistido por Computadora (Computer-Aided Diagnosis, CAD)

En 1998, La Administración de Alimentos y Drogas de los estados Unidos (U.S. Food and Drug Administration, FDA) aprobó un dispositivo que apoyaba a los radiólogos a obtener e interpretar mamografís. Este dispositivo funcionaba con un escáner, que convertía las mamografías convencionales en archivos digitales que podían ser analizados por medio de software computacional y los resultados se mostraban en una pantalla. El software que se utilizó analizaba la representación digital de las mamografías y marcaba las áreas que consideraba sospechosas para que el radiólogo emitiera un diagnostico con base en sus conocimientos, como se hacía en manera convencional de diagnostico con base en mamografías. Estudios preliminares mostraron que el software había ayudado a la detección del cáncer de mama, lo que incentivó a los involucrados en el área de la salud a invertir en el desarrollo de las tecnologías CAD y se estima que cerca del 30 % de las mamografías son actualmente interpretadas utilizando algún tipo de dispositivo CAD [43].

2.1.7. Mamografías

Una mamografía o mamograma es una radiografía del seno. El mamograma de detección se usa para encontrar enfermedades de los senos en mujeres que no tienen síntomas, es decir, que aparentemente no tienen problemas en los senos. Por lo general, en las *mamografías de detección* se toman dos radiografías (tomadas de ángulos diferentes) de cada mama.

Se usa *mamografía de diagnóstico* para especificar alguna enfermedad del seno en mujeres que presentan síntomas en sus mamas (como una protuberancia o secreción del pezón) o resultados anormales en un mamografía de detección. Un mamograma de diagnóstico incluye más imágenes del área que llama la atención del médico especialista. En algunos casos se usan imágenes especiales conocidas como vistas cónicas o de detección con magnificación para facilitar la evaluación de un área pequeña de tejido anormal del seno.

Una mamografía de diagnóstico puede mostrar:

- Que la anomalía no es motivo de preocupación alguna. En estos casos, la mujer puede volver a hacerse mamografías rutinarias cada año, lo cual es altamente recomendado.

- Que una lesión (área de tejido anormal) tiene una alta probabilidad de ser benigna (no cancerosa). En estos casos, es común que el médico indique a la paciente que debe regresar más pronto de lo usual para su próximo mamograma, generalmente en 4 a 6 meses.

- Que la lesión es motivo de más sospecha y que es necesario realizar otro tipo de examen, como una biopsia, para determinar si es cancerosa.

De acuerdo con Poggi y Harney [41] la mamografía es el medio más confiable para detectar el cáncer de mama antes que el tumor que lo causa sea palpable. En algunos casos el tumor puede ser identificado por medio de la mamografía dos años antes que este alcance un tamaño palpable.

Las mamografías se han utilizado en los Estados Unidos desde la década de 1960 y las técnicas utilizadas siguen siendo modificadas y mejoradas en la actualidad. Una mamografía convencional emite una dosis de radiación de 0.1 cGy para el estudio, y se ha comprobado que esa dosis de radiación no incrementa el riesgo de padecer cáncer de mama [31].

El Instituto Nacional del Cancer de los Estados Unidos recomienda que las mujeres de 40 años o más se deben hacer mamografías cada 1 ó 2 años y que aquellas mujeres que tienen un riesgo mayor que el promedio de presentar cáncer de mama deben hablar con un profesional médico sobre la necesidad de hacerse mamografías antes de los 40 años y sobre la frecuencia [43].

Proyeccciones de la mamamografía

De acuerdo a Kopans [34] las proyecciones de la mamografía son:

Proyección Medio-Lateral Oblicua (MLO): Es la vista de arriba hacia el centro de la mama con un ángulo de 45° desde el centro del pecho (ver Figura 2.4b).

Proyección Cráneo-Caudal (CC): Es la vista de arriba hacia abajo de la mama (ver Figura 2.4c).

Proyección Medio-Lateral (ML): Es la vista desde el centro del pecho hacia el lado exterior de la mama.

Proyección Infero-Superior Oblicua (ISO): Es la vista de abajo hacia el centro de la mama con un ángulo de 45° desde el centro del pecho.

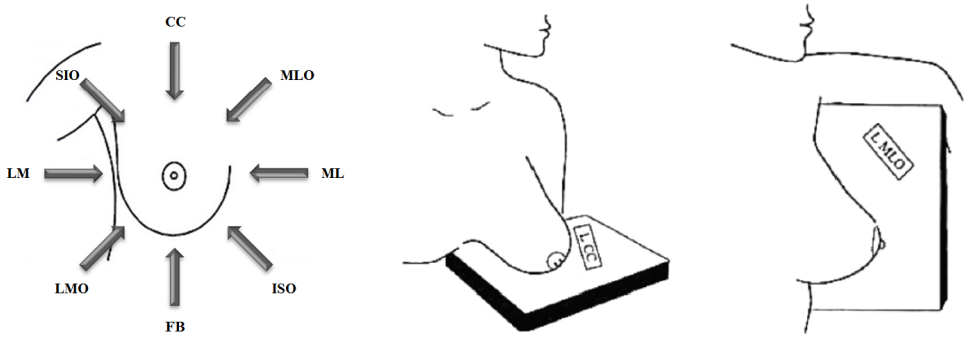
Proyección Desde Abajo o From Below(FB): Es la vista de abajo hacia arriba de la mama.

Proyección Latero-Medial Oblicua (LMO): Es la vista de abajo hacia el centro de la mama con un ángulo de 45 desde el lado exterior de la mama.

Proyección Latero-Medial (LM): Es la vista desde el lado exterior de la mama hacia el centro del pecho.

Proyección Supero-Inferior Oblicua (SIO): Es la vista de arriba hacia el centro de la mama con un ángulo de 45 desde el lado exterior de la mama.

La figura 2.3(a) muestra los ángulos de la totalidad de las proyecciones descritas arriba. La figura 2.3(b) y 2.3(c) son dos de las proyecciones con mayor uso.



(a) Proyecciones (b) CC (c) MLO

Figura 2.3: Proyecciones de la mamografía

Categorización BI-RADS de hallazgos

El Colegio Estadounidense de Radiología (American College of Radiology, ACR) ha establecido un método uniforme para que los radiólogos describan los resultados de las mamografías. El sistema, llamado BI-RADS (Breast Imaging Reporting and Database System), consiste en siete categorías o grados generalizados. Cada categoría del BI-RADS tiene un plan de seguimiento asociado para ayudar a los radiólogos y a otros médicos a manejar adecuadamente el cuidado de las pacientes.

Las categorías determinadas por el BI-RADS [19, 13] según el grado de sospecha de hallazgos son (ver cuadro 2.5):

Categoría 0: Evaluación adicional. Se necesita evaluar imágenes adicionales y/o realizar comparación con mamografías previas.

Categoría 1: Negativa. Existe simetría, sin presencia de masas, distorsiones arquitecturales o calcificaciones.

Categoría 2: Benigna. Existen lesiones de naturaleza benigna, no sospechosas de ser cáncer.

Categoría	Evaluación	Seguimiento
0	Requiere una evaluación adicional con imágenes	Requiere más estudios con imágenes para poder asignar una categoría
1	Negativo	Continuar las mamografías regulares de detección (para mujeres mayores de 40 años de edad)
2	Resultado benigno (no canceroso)	Continuar las mamografías regulares de detección (para mujeres mayores de 40 años de edad)
3	Probablemente benigno	Hacerse una mamografía de seguimiento a los seis meses
4	Anomalía sospechosa	Puede requerir biopsia
5	Muy probable que sea maligno (cáncer)	Requiere biopsia
6	Malignidad reconocida, comprobada por biopsia (cáncer)	Biopsia confirma la presencia de cáncer antes de iniciar tratamiento

Cuadro 2.5: Categorías determinadas por el BI-RADS

Categoría 3: Probablemente benigna. Existen anormalidades menores probablemente benignas. La imagen representativa es una masa con contornos regulares, sólida y no calcificada.

Categoría 4: Anormalidad sospechosa. La apariencia de la lesión no es característica de malignidad, pero la probabilidad de malignidad es susuficientemente alta. Las imágenes representativas son masas de contornos no denidos, polilobuladas y a veces mal visualizadas, de estructura heterogénea y con calcificaciones heterogéneas (amorfas o granulares).

Categoría 5: Altamente sugestiva de malignidad. Alta probabilidad de malignidad. Las imágenes representativas son masas de contornos irregulares y espiculados y calcificaciones irregulares con disposición lineal, ductal o arboriforme.

Categoría 6: Malignidad probada. Lesión identificada como tumor maligno por pruebas de biopsia.

2.1.8. Mamografía Digital

La mamografía digital es una variante de la mamografía convencional, la cual registra la imagen directamente en una lámina de película. La mamografía digital, hace una imagen electrónica del seno y la almacena como archivo de computadora. Esta información digital puede mejorarse, ampliarse o manipularse para realizar una evaluación ulterior con más facilidad que la información almacenada en película. La mamografía digital permite que el médico ajuste, guarde y recupere electrónicamente las imágenes digitales. Algunas ventajas que ofrece la mamografía digital sobre la convencional son [43]:

- Los profesionales médicos pueden compartir imágenes archivadas electrónicamente, lo cual facilita las consultas a larga distancia entre radiólogos y cirujanos.
- Es posible notar con mayor facilidad diferencias sutiles entre los tejidos normales y los anormales.
- El número de procedimientos de seguimiento necesarios puede ser menor.
- Posibilidad de repetir menos imágenes, lo cual reduce el tiempo de exposición a la radiación.

2.1.9. Elementos que se pueden observar en una mamografía

El cuadro 2.6 contiene la lista de elementos que se pueden en observar en ambas variantes de mamografías y las características de cada elemento [43].

Problemas	Características
Quistes	Son bolas llenas de líquido. Normalmente no son cáncer. Ocurren con más frecuencia en mujeres entre 35 y 50 años. A menudo ocurren en ambos senos. Algunos son demasiado pequeños y por eso no se pueden sentir al tocar los senos.
Fibroadenoma	Es una masa dura, redonda y benigna. Se siente como de goma y se mueve con facilidad. Normalmente no duele. A menudo la mujer se lo puede encontrar ella misma. Aparece en la mamografía como una bola lisa y redonda, con bordes claramente definidos. Puede crecer cuando una mujer está embarazada o dando pecho.
Macrocalcificaciones	En la mamografía aparecen como grandes acumulaciones de calcio. A menudo son causadas por el envejecimiento. Normalmente no son cáncer. Si se encuentran agrupadas de cierta manera, pueden ser un signo de cáncer.
Masas	Puede ser redonda y lisa. Puede ser causada por cambios hormonales normales. Los bordes irregulares pueden ser un signo de cáncer.

Cuadro 2.6: Elementos que se pueden ver con la mamografía

2.1.10. Masas

Una masa es definida como una lesión ocupante de espacio en tres dimensiones que es visible en al menos dos proyecciones ortogonales. Si sólo se visualiza en una proyección, debe denominarse densidad hasta comprobar su tridimensionalidad. Se describe por su posible forma, sus márgenes, densidad, localización y tamaño [19, 34, 52].

Forma de una masa

Por su forma, una masa se clasifica en:

Redonda: Masa de forma casi circular (ver Figura 2.4(a)). Este tipo de lesiones normalmente tienen una naturaleza benigna.

Ovalada: Masa de forma elíptica (ver Figura 2.4(b)), como un tipo de circunferencia alargada. Al igual que el las masas con forma redonda la naturaleza de las masas ovalada es normalmente benigna.

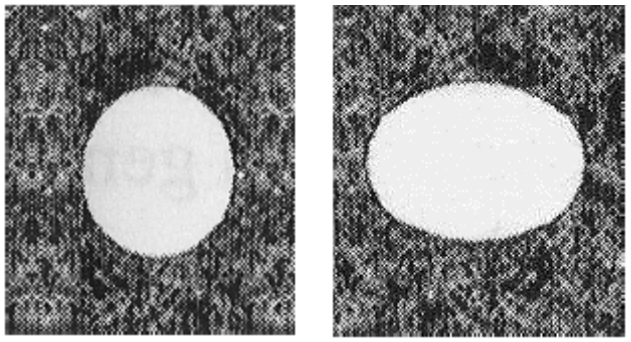
Lobular: Masa con salientes redondeadas (ver Figura 2.4(c)). Este tipo de lesión tiene una mayor probabilidad que las anteriores de ser maligna.

Irregular: Masa sin forma geométrica normal asociable (ver Figura 2.4(d)). La probabilidad de ser una lesión maligna aumenta cuanto más irregular es la forma de la lesión.

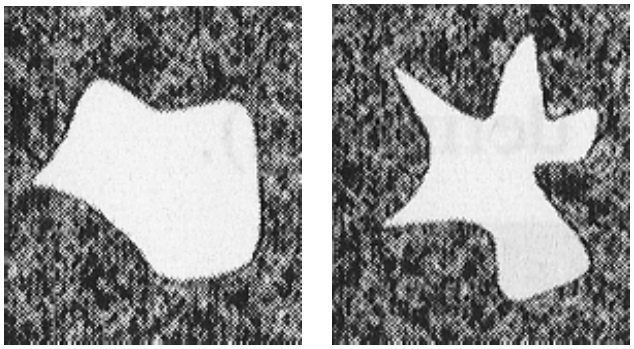
Margen de una masa

El margen que rodea a la masa es el borde entre la lesión y el tejido circundante, y es una de las características más importantes para determinar su naturaleza. Los márgenes se clasifican en [19, 52]:

Circunscrito: Margen bien denido o perfectamente denido (ver Figura 2.5(a)). Por lo general de naturaleza benigna, si existe una sola masa sólida circunscrita, se debe dar seguimiento cada 6 meses para determinar si es estable y no crece, de ser así, el seguimiento continúa cada 2 años. Para el caso de múltiples masas circunscritas, la probabilidad de benignidad aumenta, ya que son considerados como quistes, broadenomas o ganglios linfáticos intramamarios benignos, y se recomienda dar seguimiento cada año. Menos del 10 % de cánceres tienen este tipo de bordes y corresponden a carcinoma ductal inltrante, carcinoma papilar o carcinoma medular.



(a) Redonda (b) Ovalada



(c) Lobular (d) Irregular

Figura 2.4: Formas de masas mamarias según clasificación BI-RADS.

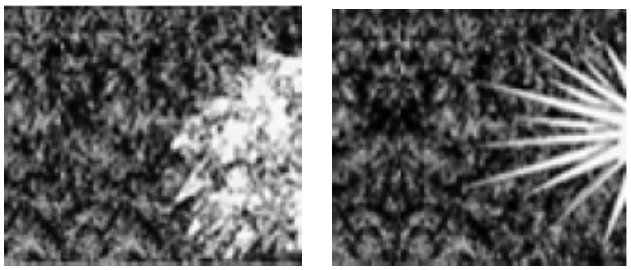
Microlobulado: Margen con pequeñas ondulaciones en su límite, semejantes a pétalos de una or (ver Figura 2.5(b)). Tiene mayor probabilidad de delimitar una masa cancerosa, sin embargo, existen hallazgos benignos con este tipo de margen, como los broadenomas y quistes.

Ensombrecido: Margen borroso que queda oculto por el tejido broglandular adyacente y por esta razón no puede ser valorado fácilmente (ver Figura 2.5(c)). De igual forma que el anterior, tiene mayor probabilidad de delimitar una masa maligna.

Mal definido: Margen que no tiene forma específica por invadir tejido circundante debido a la infiltración de las células cancerosas en la periferia de la masa y no puede atribuirse al tejido normal superpuesto (ver Figura 2.5(d)). Tiene alta probabilidad de delimitar una masa maligna, sin embargo y similar al margen microlobulado, hallazgos benignos con este margen son los broadenomas y quistes.

Espiculado: Margen que se caracteriza por sus finas líneas o patas que se proyectan de forma radial desde la zona central de la masa hacia su periferia (ver Figura 2.5(e)). Tiene muy alta probabilidad de albergar un tumor de naturaleza maligna. La mayoría de hallazgos con este tipo de margen corresponden a un carcinoma ductal infiltrante y pocos casos corresponden a carcinomas tubulares y lobulillares.

(a) Circunscrito (b) Microlobulado (c) Ensombrecido



(d) Indefinido (e) Espiculado

Figura 2.5: Margen de masas mamarias según clasificación BI-RADS.

2.2. Aprendizaje automático

Para poder resolver un problema utilizando una computadora, es necesario un algoritmo. Un algoritmo es una secuencia de instrucciones por ejecutar de tal forma que permiten convertir una entrada en una salida especifica. En la actualidad, existen algoritmos para resolver una gran cantidad de problemas diversos, así como también existen aún muchos problemas sin resolver. Esos problemas sin resolver se han convertido en objeto de múltiples investigaciones al grado que hay personas dedican su vida buscar la solución a dichos problemas. Para algunos problemas, si bien no se ha podido desarrollar una solución, se han podido desarrollar buenas aproximaciones, capaces de ser utilizadas con si se tratase de la solución misma [2]. El aprendizaje automático es un sinónimo de computación avanzada; computación donde se han desarrollado técnicas que permiten resolver problemas del mundo real. La compleja o no entendida naturaleza de muchos problemas ha creado la necesidad de desarrollar tecnologías que puedan adaptarse a la tarea que se les encomienda realizar [8]. En este sentido el objetivo del aprendizaje automático es que los sistemas puedan cambiar su comportamiento de manera autónoma basados en la experiencia, entendiéndose como esta a la información histórica la que tiene acceso el sistema y que puede ser utilizada por el mismo en su labor. A este proceso de utilizar la experiencia para producir conocimiento se le conoce como aprendizaje [42].

De forma general, el aprendizaje automático inicia con un conjunto de entradas a las que se les denomina *ejemplos*, los cuales están compuestos de diferentes *características* y que puede tener asociada un *clase* o *etiqueta de clase*.

2.2.1. Aplicación y problemas

Los algoritmos de aprendizaje automático han sido ampliamente utilizados en aplicaciones diversas como:

- Clasificación de textos y documentos.
- Procesamiento del lenguaje natural.
- Reconocimiento de voz y verificación de identidad por voz.
- Reconocimiento de imágenes y detección de rostros.
- Sistemas de recomendación, motores de búsqueda y sistemas de extracción de información.
- Diagnóstico médico.

Los problemas de aprendizaje se dividen, principalmente en las siguiente áreas:

- **Ranqueo:** El ranqueo consiste en determinar una relación entre un conjunto de elementos tales que, de acuerdo a uno o varios criterios, se puede definir un orden entre los miembros del conjuntos de tal forma que el primero de ellos presenta un valor superior al segundo, este a su vez mayor que el tercero y así sucesivamente [42].

- **Agrupamiento:** En esta área del aprendizaje automático se tiene como tarea el realizar una partición del conjunto de entradas en grupos, excluyentes entre si, y de tal forma que los miembros de una agrupación (denominada *cluster*) sean similares entre si y deferentes a los miembros de otras agrupaciones [42].

- **Reducción de dimensiones:** Proceso mediante el cual se pretende transformar un conjunto de entradas X_i de la forma $x_1, x_2, ..., x_n$ a un espacio con menores dimensiones, conservando ciertas propiedades de la representación original [2, 42].

- **Regresión:** Se denomina *regresión* al aprendizaje automático donde las clases a predecir y son números reales.Un ejemplo de regresión sería un sistema que puede predecir el precio de un carro usado. La entrada de este sistema serían los diferentes atributos del auto (marca, color, modelo, etc.). Estos elementos que componen la entrada están directamente relacionados al valor buscado, el importe de dicho auto. En este ejemplo, como la salida esperada del sistema es un número entero entonces, se dice que el problema es de regresión. [2, 42]

- **Clasificación:** Tipo de aprendizaje automático donde a partir de los elementos X_i del conjunto de entrenamiento del sistema y su respectiva etiquetas de clase y_i se aprende a realizar el mapeo de entradas a salidas. El conjunto de elementos y_i es finito, con elementos discretos y normalmente tiene baja cardinalidad. En términos generales, la clasificación hace referencia al problema de identificar a que elemento de un conjunto de categorías pertenece una determinada entrada.Un ejemplo de problema de clasificación es el que tienen algunas instituciones financieras cuando un persona solicita un crédito. Un crédito es una cantidad de dinero que una institución bancaria otorga a una persona o a otra institución, la cual será pagada en un futuro agregando un cierto porcentaje de interés. La labor de la institución financiera consiste en predecir el riesgo asociado a una determinada solicitud de préstamo, el cual se puede ver como la probabilidad que el prestatario devuelva el monto prestado al prestamista. En este sentido, la institución prestamista debe calcular el riesgo de una solicitud con base en la cantidad requerida y la información del prestatario. Este es un ejemplo de clasificación, porque el banco debe asignar a los clientes en categorías, como podrían ser clientes de bajo y alto riesgo. La información sobre un nuevo prestatario constituye la entrada del *clasificador*, y éste tiene como labor decidir cual de las dos categorías de clientes corresponde a la estrada en cuestión. El algoritmo de aprendizaje adquiere la capacidad de decidir la clase de un cliente a partir de la historia de clientes y prestamos con la

que cuenta la institución en cuestión. La meta de la clasificación, al igual que la regresión, es construir un modelo capaz asignar a un patrón la clase correcta a la que corresponde. En el aprendizaje automático, se presenta a un algoritmo *clasificador* un conjunto de patrones para que dicho algoritmo *aprenda* a realizar dicho mapeo. La entrada de un algoritmo de clasificación es un conjunto de patrones, los cuales se constituyen de atributos. Uno de esos atributos es dependiente de los demás y cuyo valor representa a la clase a la que pertenece un patrón. El resto de atributos independientes son utilizados por un *clasificador* para asignar un patrón a la correspondiente clase [2, 40, 42].

2.2.2. Esquemas de aprendizaje automático

El Aprendizaje Automático es la rama de la Inteligencia Artificial que tiene como objetivo desarrollar técnicas que permitan a las computadoras aprender a partir de datos o experiencia previa [2]. Para llevar a cabo el aprendizaje se utiliza un conjunto de datos de *entrenamiento* el cual es un conjunto de vectores X_i de la forma $x_1, x_2, ..., x_n$, donde los x_j representan las característica o atributos que en combinación permiten diferenciar entre vectores de la misma naturaleza.

Los algoritmos de aprendizaje automático utilizan una función $y(X_i)$ que recibe como entrada un vector de características y regresa como salida una etiqueta de clase y. Esta función $y(X_i)$ es determinada en la etapa de entrenamiento o de aprendizaje y tiene como base la información que aporta el conjunto de datos X_i. Una vez que un algoritmo de aprendizaje es entrenado, éste es capaz de realizar predicciones de nuevos ejemplos X_k y el resultado de dicha predicción depende del modelo interno generado en la fase de entrenamiento. La habilidad de un algoritmo para realizar un *correcta* asociación entre ejemplos X_k y su correspondiente clase y se le denomina *generalización*.

Con base en la forma de utilizar la información histórica de la que se dispone para el aprendizaje automático, se consideran principalmente en cuatro enfoques:

- **Supervisado:** En este tipo de aprendizaje, los sistemas deben aprender a realizar un mapeo entre las entradas X_i y salidas y a partir del conjunto de datos de entrenamiento y su respectiva salida, de la cual el sistema dispone de manera *a priori*.

- **No-Supervisado:** En este tipo de problemas solamente se tiene acceso a las entradas X_i y el algoritmo debe encontrado similitudes entre dichos datos que permitan producir las salidas y.

- **Semi-supervisado:** En la práctica, resulta complicado poder contar con un conjunto completo de ejemplos X_i y sus respectivas salidas y. A veces solamente una parte de los X_i cuentan con su correspondiente y, en ese caso se dice que el aprendizaje es semi-supervisado.

- **Por refuerzo:** En este tipo de aprendizaje el algoritmo tiene como finalidad obtener la recompensa máxima que se recibe por las acciones que se realiza. Una acción que contribuye a la correcta solución recibe una recompensa, mientras que a una que no se le otorga un castigo.

2.2.3. Entrenamiento

En aprendizaje automático se entiende como entrenamiento a la fase mediante la cual un algoritmo utiliza un conjunto de ejemplos, a los que se les denomina como *conjunto de entrenamiento* para desarrollar un modelo interno que permita general un mapeo de entradas salidas. De acuerdo a la forma en que se realiza el entrenamiento, se pueden distinguir donde enfoques principales de aprendizaje automático [42]:

- **Fuera de línea (Off-line training):** En este escenario de aprendizaje el algoritmo en cuestión recibe la totalidad de ejemplos de una vez. En este esquema de entrenamiento el algoritmo de aprendizaje mantiene el modelo interno desarrollado a partir de los ejemplos utilizados en el entrenamiento.
- **En línea (On-line training):** A diferencia del esquema fuera de línea, el aprendizaje en línea realiza una actualización periódica del modelo interno del algoritmo de aprendizaje.

2.3. Reducción de Dimensiones

Uno de los factores que influyen directamente en la complejidad de un clasificador es el número de entradas que este recibe. Estas entradas determinan la complejidad tanto en tiempo de ejecución como en memoria requerida. En [2, 15, 16, 66] se exponen algunas razones por las cuales se recomienda realizar un proceso de reducción de las entradas de los problemas en aprendizaje automático, a continuación se listan algunas de las más importantes:

- En la mayoría de los algoritmos de aprendizaje, la complejidad depende de las dimensiones de las entradas. En este sentido se puede decir que la complejidad y las dimensiones de las entradas guardan una relación de monotonía.
- Cuando se puede decidir que una entrada es innecesaria, se ahora el tiempo de computo que requería procesas dicha entrada.
- Modelos simples son más robustos y sufren menos variaciones en conjuntos de datos pequeños.
- Cuando los datos pueden ser explicados con pocos elementos, es más fácil comprender el proceso que produce la salida.

Existen diferentes técnicas que se emplean para descubrir relación entre las características del dominio de un problema. Algunas tienen su base en el análisis estadístico como es el caso del análisis de *correlación*. Otra manera de realizar la reducción de dimensiones es por medio de un algoritmo de *búsqueda* que explore el espacio de subconjuntos posibles con el objeto de encontrar el conjunto de menor tamaño con rendimiento igual o mejor al conjunto original de características. A continuación se presentan tres enfoques comúnmente utilizados para la reducción de dimensiones:

2.3.1. Análisis de correlación

Se utiliza el método de la matriz de la correlación para encontrar si existe relación alguna en cuanto a variables involucradas en el dominio de un problema [15, 45].

Si dos características evolucionan de modo tal que en alguna medida se relacionan entre ellas, podemos decir que existe una asociación entre ellas. La asociación entre las características no significa que una de ellas dependa causalmente de la otra, puede ser una pura coincidencia. Una forma de expresar la asociación entre dos características es la correlación del momento-producto o correlación de Pearson. Si el coeficiente de correlación es bajo (entre $-0{,}3$ y $+0{,}3$) las dos características no están asociadas entre sí. Si es alto (cercano a $+1$ o a -1) significa que la relación entre ellas se aproxima a la ecuación $y = Ax + B$. El signo del coeficiente de correlación no es importante (es el signo de A en la ecuación de la recta). Un aspecto débil del análisis de correlación es que no puede detectar otras relaciones no lineales entre las características, por ejemplo $y = Ax^2 + Bx + C$ pasaría inadvertida. Luego de detectar que la correlación entre dos características es bastante alta se procede a eliminar una de ellas. Para ello se calcula la ganancia de información, y se descarta a aquella que aporta menor poder para discriminar.

2.3.2. Discretización de características

Para calcular la ganancia de información, es necesario trabajar con variables (en este caso características) discretas. Un algoritmo de discretización permite dividir los valores de las variables en particiones de intervalos discretos. Las técnicas de discretización se clasifican en:

1. Supervisadas y no supervisadas: Las discretización supervisada usa la información de la clase a la que pertenece cada variable, las segundas intentan identificar la clase a la que pertenecen desde los datos proporcionados.

2. Globales y locales: Los primeros intentan discretizar todos los rangos de la variable, los segundos intentan hacerlo por rangos.

Existen diversas técnicas de discretización, entre las que se resaltan:

1. Intervalos de igual anchura: Permite dividir los datos en k intervalos de igual anchura. La técnica es no supervisada, el valor apropiado de k se debe de calcular por experimentación y es muy sensible a los valores extremos.

2. Intervalos de igual frecuencia: Se dividen los datos en intervalos de igual frecuencia. El método es no supervisado.

3. Particiones de mínima entropía: Intenta identificar el mejor umbral para dividir los datos en intervalos de tal forma que la *entropía* de cada uno sea la menor posible.

El pseudocódigo 2.1 muestra el proceso para realizar la discretización no supervisada y global, buscando intervalos de igual anchura.

Algoritmo 2.1 Discretización

Entrada: $S_m, intervalos$

1: **desde** $i = 1$ **hasta** $i = columnas(S_m)$ **hacer**
2: data $= S_m[i]$
3: Max $= max$(data)
4: Min $= mix$(data)
5: paso $= \frac{Max-Min}{intervalos}$
6: **desde** $j = 1$ **hasta** $j = filas(S_m)$ **hacer**
7: discreto$[j] = entero(\frac{data[j]}{paso})$
8: **fin desde**$S_d[i]$ = discreto
9: **fin desde**
10: **devolver** S_d

Se reciben como entradas las características obtenidas de para cada masa tumoral S_m. Posteriormente, se obtienen como resultado las características discretizadas en intervalos S_d. El algoritmo recibe como entrada el número de intervalos en que se divide el rango de valores, *intervalos*. La información contenida en S_d será utilizada para calcular la *ganancia de información*.

2.3.3. Ganancia de información

Este procedimiento intenta aprovechar el concepto de ganancia de información usado en la construcción de árboles de decisión para determinar qué características extraídas desde las lesiones tienen la mayor ganancia de información individual para predecir la clase a la que pertenece [35, 45]. Se calcula la ganancia de información de cada característica por separado y luego se les ordena de mayor a menor. La ganancia de información se puede calcular desde la *entropía*.

Entropía

La entropía mide el grado de incertidumbre que se asocia a una distribución de probabilidad. Por ejemplo, en una distribución uniforme, todos los valores son igualmente probables $p_i = \frac{1}{N}$ y por tanto la entropía es máxima, lo cual indica máxima incertidumbre. Por el contrario, en una distribución con un solo pico en la que $p_i = 1$ y $p_j = 0$, para todo $j \neq i$ la entropía es mínima lo cual indica mínima incertidumbre.

La ecuación 2.1 muestra la forma de calcular la entropía de una variable discreta S:

$$E(S) = \sum_{i \in C} -p_i \log_2 p_i \tag{2.1}$$

Donde S es un conjunto que se puede dividir en $|C|$ clases, pi es la proporción de ocurrencias de la clase i en el conjunto S, de la siguiente forma:

$$p_i = \frac{|S_i|}{S} \tag{2.2}$$

Ganancia de información

La ganancia de información se usa frecuentemente para construir árboles de decisión, permite decidir qué característica A adicionar al árbol actual. Es una medida de cuánto ayuda el conocer el valor de cierta característica para conocer el verdadero valor de la clase a la que pertenece la masa asociada. Una ganancia de información alta implica que una característica permite reducir la incertidumbre de la clase a la que pertenece la lesión.

La ganancia de información se puede calcular a partir de la entropía global y la media ponderada de las entropías asociadas a los valores que puede tomar una característica. Para calcular la ganancia asociada a cada característica A, se usa la siguiente fórmula:

$$ganancia(S, A) = entropía(S) - \sum_{\forall v, v \in A} \frac{|S_v|}{S} entropía(S_v) \tag{2.3}$$

Donde S_v, es un subconjunto de S, donde la característica A toma el valor de v.

El pseudocódigo 2.2 muestra el proceso para realizar la discretización no supervisada y global, buscando intervalos de igual anchura. El procedimiento recibe como entradas los valores discretizados de cada característica S_d. Se obtiene como resultado la lista de características ordenadas en función a su ganancia de información S_i. Tiene un parámetro, la clase a la que pertenece cada masas, *clase*.

Algoritmo 2.2 Ganancia de Información

Entrada: $S_d, clase$

1: $E_S = 0$ entropía de la clase
2: $E_{SA} = 0$ media ponderada de la entropía
3: **desde** $i = 1$ **hasta** $clases(clase)$ **hacer**
4: $\quad p_i = \frac{tamaño(clase(i))}{tamaño(clase)}$
5: $\quad E_s = E_s - p_i \log_2 p_i$
6: **fin desde**
7: **desde** $i = 1$ **hasta** $tamaño(S_d)$ **hacer**
8: $\quad A = S_d[i]$
9: $\quad$ **desde** $v = 1$ **hasta** $clases(A)$ **hacer**
10: $\quad\quad E_v = 0$
11: $\quad\quad$ **desde** $c = 1$ **hasta** $clases(clase)$ **hacer**
12: $\quad\quad\quad p_i = \frac{tamaño(A(v,c))}{tamaño(A(c))}$
13: $\quad\quad\quad E_v = E_v - p_i \log_2 p_i$
14: $\quad\quad$ **fin desde**
15: $\quad\quad E_{SA} = E_{SA} + \frac{tamaño(clase(v))}{tamaño(clase)} E_v$
16: $\quad$ **fin desde**
17: $\quad ganancia[i] = E_S - E_{SA}$
18: **fin desde**
19: **devolver** $S_i = ordenar(S_d, ganancia)$

2.3.4. Búsqueda secuencial

El algoritmo posiblemente más sencillo que se podría idear para resolver un problema de optimización ciega consiste en ir evaluando todos los puntos en el espacio de búsqueda de alguna manera ordenada que evite repeticiones, es decir, hacer intentos de acuerdo a una regla que asegure que, si el algoritmo se corre por suficiente número de intentos, cada punto del espacio de búsqueda será evaluado una vez, y solamente una vez. A este algoritmo se le denomina búsqueda secuencial.

El pseudocódigo 2.3 implementa la búsqueda secuencial:

Algoritmo 2.3 Búsqueda secuencial

1: $x \leftarrow$ generar primer elemento de la secuencia
2: **repetir**
3: $\quad x_s \leftarrow$ generar siguiente elemento de la secuencia
4: $\quad$ **si** x_s es mejor que x **entonces**
5: $\quad\quad x \leftarrow x_s$
6: $\quad$ **fin si**
7: **hasta** cumplir criterio de terminación
8: **devolver** x

El criterio de terminación que aparece en este código depende de la aplicación y del usuario. Algunas maneras usuales de terminar un algoritmo de optimización ciega son mediante el número de intentos, tiempo de ejecución, calidad de la solución alcanzada, y número de intentos sin encontrar una solución mejor a la actual.

2.3.5. Algoritmos genéticos

Los *algoritmos genéticos* y recocido simulado son métodos de optimización ciega. Por optimización ciega se entiende la búsqueda de los parámetros que hacen que se optimice (maximice o minimice) la evaluación de una función dada. En optimización ciega no se asume que se tenga acceso a la función que se está optimizando, únicamente es necesario tener acceso a su evaluación.

Las técnicas de optimización ciega son de gran importancia práctica debido a que con ellas muchos problemas que inicialmente no serían concebidos como problemas de optimización, pueden ser atacados como problemas de optimización. Como algoritmos de búsqueda, que recorren un espacio de soluciones posible en busca de la mejor, pueden ser adaptados para resolver problemas muy diversos.

Los algoritmos genéticos son una técnica que puede utilizarse en la etapa de reducción de dimensiones. Un algoritmo genético es una técnica que trabaja sobre una población de individuos mediante operadores de selección, cruce, mutación y reemplazo evoluciona hasta encontrar el mejor individuo de l población. La inspiración de este tipo de algoritmo es la teoría de la evolución de las especies y las mejoras y adaptación que debe de tener una especie para sobrevivir con el paso de tiempo. Utiliza una función de evaluación, que permite que se cumpla la analogía de la supervivencia del más apto [4].

Un algoritmo genético (AG) un método estocástico de búsqueda basado en la mecánica de selección natural y la idea darwiniana de la supervivencia de acuerdo a la aptitud. Se utilizan tiras de caracteres de un alfabeto reducido, por lo general binario, para representar las soluciones o individuos de un problema, semejante a los cromosomas de los seres vivos. Cada población de soluciones evoluciona al aplicar operadores genéticos como selección, cruce y mutación [22].

Para implementar un algoritmo genético [61] se debe primero determinar la función de aptitud y la representación de las soluciones. Una solución es, por lo general, representada por un arreglo de bits, a los que se les denomina genes, de tamaño fijo o alguna estructura de uso similar. El valor que toma cada bit en la representación de los individuos, representa las diferencias genéticas existente entre miembros de una misma especie. A partir de estas diferencias y mediante una función de aptitud, se puede determinar que un individuo es mejor otro en una población dada. La función de aptitud (o *fitness*) mide la calidad de la solución representada por un individuo.

El código 2.4 muestra el funcionamiento de una algoritmo genético:

Algoritmo 2.4 Algoritmo Genético

Entrada: Tamaño de población n, Número de generaciones N
Salida: Solución mejor encontrada $mejor$

1: $poblacion = generar\ poblacion\ aleatoria(n)$
2: $aptitudes, mejor = evaluar\ aptitudes(poblacion)$
3: **repetir**
4: $g = g + 1$ Contador de generaciones
5: $poblacion = aplicar\ seleccion(poblacion, aptitudes)$
6: $poblacion = aplicar\ cruce(poblacion)$
7: $poblacion = aplicar\ mutacion(poblacion)$
8: $aptitudes, mejor = evaluar\ aptitudes(poblacion)$
9: **hasta** cumplir criterio de terminación
10: **devolver** $mejor$

Los puntos más importantes para resolver un problema con algoritmos genéticos consisten en realizar un correcta representación de las soluciones, a manera de simplificar la aplicación de los operadores cruce, mutación y selección y poder acceder a una función de aptitud para poder evaluar a los individuos de l población.

2.3.6. Metodología de envoltura (Wrapper)

La metodología de la *envoltura (wrapper)*, es bastante popular y ofrece un manera simple y poderosa de atacar problemas de reducción de dimensiones y selección de características, sin importan el algoritmo de clasificación con el que se esté trabajando [24].

En este enfoque, el algoritmo de clasificación se usa como parte de la estrategia búsqueda dentro del espacio de características. Las secciones anteriores muestran como existen diferentes enfoques de búsqueda que permiten explorar el espacio de posibles subconjuntos de características en busca de aquel que contenga el menor de numero de atributos y que aporte tenga el mejor desempeño. En el modelo de la envoltura la evaluación de conjunto de características se hace mediante la construcción y entrenamiento de algún algoritmo de clasificación, del cual se obtiene diferentes métricas de desempeño, siendo estas asignadas al subconjunto de características en cuestión. Se dice que el algoritmo de búsqueda *envuelve* al algoritmo de clasificación, debido a que este último es tratado como caja negra para obtener las métricas de desempeño de ese subconjunto.

La figura 2.6 muestra un bosquejo del modelo de la envoltura.

La metodología de la envoltura es comúnmente criticada, bajo la idea que se trata de un algoritmo de "fuerza bruta"que requiere un gran esfuerzo computacional. Todo lo anterior puede ser descartado si se usa un buen algoritmo de búsqueda, como los algoritmo genéticos.

El modelo de selección de características que fue estudiado en la presente investigación corresponde al modelo de *wrapper* y se utilizo como algoritmo de búsqueda un

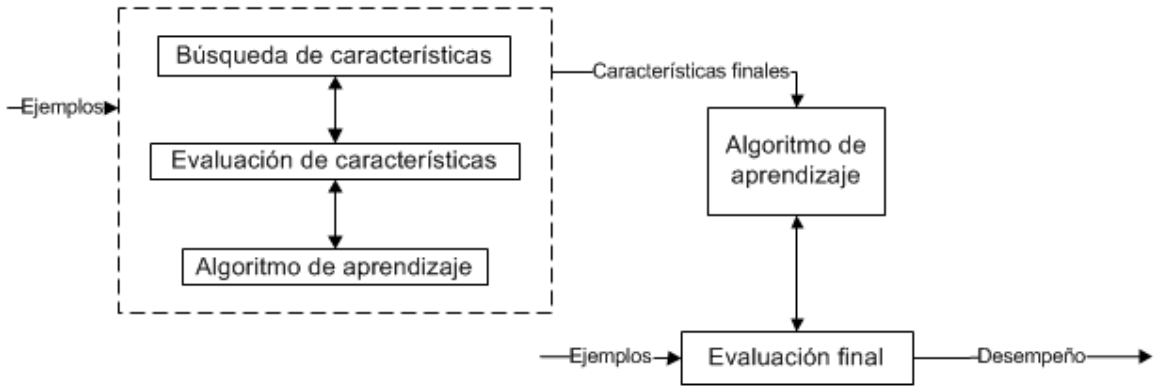


Figura 2.6: Metodología de envoltura (Wrapper)

algoritmo genético y cuatro algoritmos de clasificación diferentes algoritmos de clasificación para evaluar las posibles soluciones.

La siguiente sección aborda los aspectos generales de los algoritmos de clasificación considerados en la presente investigación.

2.4. Algoritmos de clasificación

Un clasificador es un algoritmo que pretende asignar un objeto a una clase determinada en función de los atributos (aquí denominados características) que se le suministran como entrada [45]. La presente sección muestra los aspectos generales de los algoritmos: Redes Neuronales (NN), Máquinas de Vector de Soporte (SVM), Análisis de Discriminante Lineal (LDA) y Regresión Logística (LR).

2.4.1. Redes Neuronales

Las redes neuronales constituyen un paradigma de aprendizaje automático inspirado en la forma en como funciona el sistema nervioso de los animales. Una red neuronal es un sistema de interconexión de elementos que colaboran entre sí para producir un estimulo de salida [4]. La red está conformada por muchos elementos computacionales (*nodos* o *neuronas*) no lineales que operan en paralelo. Los nodos están conectados en *capas* y las *conexiones* entre nodos tienen *pesos*, que se van modificando durante el proceso de entrenamiento.

El algoritmo de *retro-propagación* es usado para entrenar los pesos de la red en dos fases, en la primera un patrón de entrenamiento es presentado propagándose a través de la red hasta la salida, donde se calcula el *error* (diferencia entre salida deseada y salida obtenida); en la segunda el error se trasmite hacia atrás, hasta los nodos de la capa de entrada, recibiendo cada nodo un porcentaje del error. Con base en el error se ajustan

los pesos de los nodos [10, 45]. El ciclo en el que se genera la salida, se calcula el error y se retro-propaga a los nodos de la red recibe el nombre de *época*.

El aprendizaje en las redes neuronales consiste en proporcionar a la red un conjunto de entradas y las salidas esperadas de la misma. De esta forma, modificando en cada ciclo los pesos de interconexión la red es ajustada, en cada época, hasta que la salida es igual a la salida esperada para una determinada entrada.

Cada elemento de la red consta de los siguientes elementos:

1. Entradas x_j: Son las salidas de otras unidades.

2. Pesos sinápticos w_{ij}: Conjunto de pesos de la conexión que sale unidad j y entra en la unidad i.

3. Función suma net_i: Suma ponderada por los pesos de las entradas:

$$net_i = \sum_j w_{ij} x_j \tag{2.4}$$

4. Función de activación a_i: Es función de la activación anterior y la entrada neta net_i

$$a(t) = F_i(a(t-1), net_i) \tag{2.5}$$

5. Una salida x_i: Que es función de la activación a_i:

$$x_i = f(a_i) \tag{2.6}$$

La figura 2.7 muestra una red neuronal de 3 capas, con n unidades en la primera capa o *capa de entrada*, m en la *capa intermedia* y 1 única neurona en la *capa de salida*. Como puede observarse el número de unidades en la capa de entrada corresponde al número de características que se le pasan como entrada a la red.

Teorema de Kolgomorov

"Dada cualquier función continua $f : [0,1]^n \rightarrow R^m$; $f(x) = y$, f puede ser implementada exactamente por una red neuronal de tres capas sin retroalimentación que tiene una capa de entrada de n elementos que únicamente copian las entradas a la siguiente capa, $2n + 1$ elementos de procesamiento en la capa intermedia, y m elementos de procesamiento en la capa de salida".

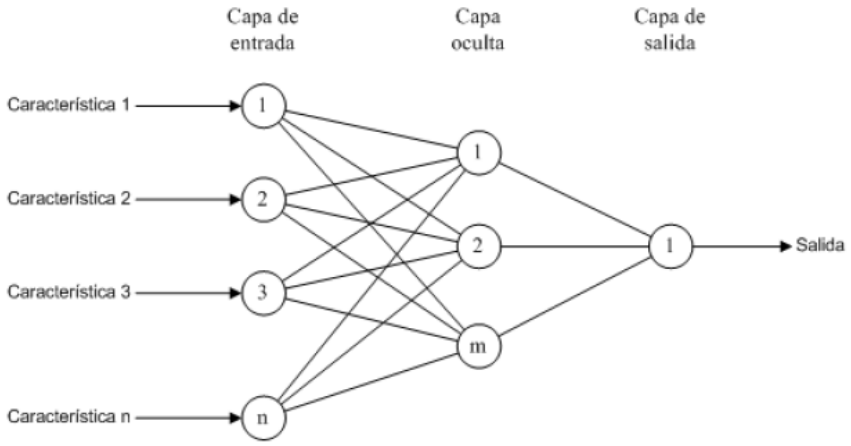


Figura 2.7: Red neuronal de 3 capas

2.4.2. Máquinas de Vector de Soporte (Support Vector Machine, SVM)

En esta sección se presenta una descripción detallada de la arquitectura de las máquinas de vector de soporte, el proceso de estimación de sus parámetros y las funciones de núcleo comúnmente utilizadas.

Una *máquina de vector de soporte, SVM* es un método de clasificación supervisada. SVM emplea un algoritmo de optimización para determinar la frontera óptima entre dos grupos, aunque se puede generalizar para múltiples grupos. El caso más simple es cuando los grupos son linealmente separable, donde existe una distancia positiva entre ambos grupos y es posible elegir un *hiperplano* de separación maximizando la distancia de éste a cada grupo. Para decidir a qué grupo pertenece una nueva observación, se considera el signo de la función que define al hiperplano de separación [55].

Los SVM, aplicados a problemas de clasificación, realizan un mapeo de los datos a un espacio de características alto dimensional, donde es posible encontrar un hiperplano de separación entre las clases. Este mapeo puede ser llevado a cabo aplicando una función tipo Kernel que transforma implícitamente el espacio de entrada en un espacio de características de mayor dimensión. El hiperplano de separación es calculado maximizando la distancia de los patrones más cercanos.

El objetivo es construir un hiperplano de decisión, de tal forma que la separación entre los objetos de una clase y de de otra es máxima. En el caso no lineal, dado un conjunto de N ejemplos y las respectivas clases a las que pertenecen $(\mathbf{X}_i, y_i)$, donde $\mathbf{X}_i \in R^n$ y $y_i \in \{-1, 1\}$, el hiperplano de separación está definido por la ecuación 2.7:

$$f(\mathbf{X}_q) = \sum_{i=1}^{N} y_i \alpha_i K(\mathbf{X}_q, \mathbf{X}_i) + b \tag{2.7}$$

En la ecuación 2.7, $K(\mathbf{X}_q, \mathbf{X}_i)$ es un función *kernel* y el signo de la función $f(\mathbf{X}_q)$

determina la clase a la que pertenece el ejemplo en cuestión $\mathbf{X}_q$. Las funciones de tipo kernel se utilizan para hacer mapeos no lineales, y los principales kernel utilizados en aprendizaje automático y reconocimiento de patrones son:

Polinomial

$$K(\mathbf{x}, \mathbf{x}_i) = (\mathbf{x} \cdot \mathbf{x}_i + 1)^d \tag{2.8}$$

Gaussian Radial Basis Function (RBF)

$$K(\mathbf{x}, \mathbf{x}_i) = e^{\frac{\|\mathbf{x}-\mathbf{x}_i\|}{2\sigma^2}} \tag{2.9}$$

Sigmoidal hyperbolic tangent

$$K(\mathbf{x}, \mathbf{x}_i) = \tanh(\kappa \mathbf{x} \cdot \mathbf{x}_i - \delta) \tag{2.10}$$

Algunas de las ventajas de los SVM van desde evitar el problema de sobre-ajuste del que pueden llegar a padecer las redes neuronales, permite trabajar con datos en espacios no lineales y se resalta su capacidad para generalizar bien incluso cuando hay pocos ejemplos disponibles.

2.4.3. Análisis Discriminante Lineal (Linear Discriminant Analysis, LDA)

El *análisis de discriminante lineal* (LDA, por sus siglas en inglés) es un método para clasificación que divide el espacio muestral en subespacios mediante hiperplanos que permiten separar los patrones de entrenamientos en clases. La idea central del LDA es obtener una proyección de los datos en un espacio de menor (incluso igual) dimensión que los datos de entrada con el fin de que la separabilidad de las clases sea la mayor posible.

En el caso de clasificación binaria, se desea decidir entre dos clases Π_1 y Π_2. Para el análisis se define que:

$$P(\mathbf{X} \in \Pi_i) = \pi_i \tag{2.11}$$

donde $i = 1, 2$ representa la probabilidad *a priori* que una observación $\mathbf{X} = \mathbf{x}$ pertenezca a Π_1 o Π_2. También se define que la probabilidad condicional que $\mathbf{X}$ pertenezca a la clase i es:

$$P(\mathbf{X} = \mathbf{x} | \mathbf{X} \in \Pi_i) = f_i(\mathbf{X}) \tag{2.12}$$

Si se aplica el *teorema de Bayes* a las ecuaciones 2.11 y 2.12 para obtener la probabilidad *a posteriori* que la observación $\mathbf{x}$ pertenezca a la clase Π_i:

$$P(\Pi_i|\mathbf{x}) = P(\mathbf{X} = \mathbf{x}|\mathbf{X} \in \Pi_i) = \frac{f_i(\mathbf{x})\pi_i}{f_1(\mathbf{x})\pi_1 + f_2(\mathbf{x})\pi_2} \tag{2.13}$$

Entonces, para un dado $\mathbf{x}$, una estrategia de clasificación es asignar $\mathbf{x}$ a la clase que obtiene una mayor probabilidad *a posteriori*. Esta estrategia es conocida como *clasificador de regla de Bayes*. De esta manera, si

$$\frac{P(\Pi_1|\mathbf{x})}{P(\Pi_2|\mathbf{x})} > 1 \tag{2.14}$$

$\mathbf{x}$ es asignado a la clase Π_1, de lo contrario a la Π_2.

Si se combinan las ecuaciones 2.13 y 2.14 se tiene que el clasificador de regla de Bayes asigna $\mathbf{x}$ a la clase Π_1 cuando

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} > \frac{\pi_2}{\pi_1} \tag{2.15}$$

de lo contrario se le asigna la clase Π_2. En el caso en que $\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} = \frac{\pi_2}{\pi_1}$ se asigna una clase de manera aleatoria.

Si se sigue lo establecido por Fisher, asumiendo que las distribuciones definidas en 2.12 son distribuciones normales con vectores arbitrarios como media, pero la misma matriz de covarianza, se puede hacer más específica el clasificador de regla de Bayes obtenido en 2.15. Esto es, si se tiene que $f_1 = N_r(\mu_1, \mathbf{\Sigma}_1)$ y $f_2 = N_r(\mu_2, \mathbf{\Sigma}_2)$ donde $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{\Sigma}_{XX}$ entonces:

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} = \frac{exp\{-\frac{1}{2}(\mathbf{x} - \mu_1)^\tau \mathbf{\Sigma}_{XX}^{-1}(\mathbf{x} - \mu_1)\}}{exp\{-\frac{1}{2}(\mathbf{x} - \mu_2)^\tau \mathbf{\Sigma}_{XX}^{-1}(\mathbf{x} - \mu_2)\}} \tag{2.16}$$

donde los factores de normalización, tanto para numerados como para denominador es $(2\pi)^{-\frac{r}{2}}|\mathbf{\Sigma}_{XX}|^{-\frac{1}{2}}$ se cancelan debido a que no aportan al cociencte. Si se aplica el logaritmo natural (base e) a ambos lados de la ecuación 2.16 se obtiene:

$$\log_e\left(\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})}\right) = (\mu_1 - \mu_2)^\tau \mathbf{\Sigma}_{XX}^{-1}\mathbf{x} - \frac{1}{2}(\mu_1 - \mu_2)^\tau \mathbf{\Sigma}_{XX}^{-1}(\mu_1 + \mu_2) \tag{2.17}$$

Y continuando con el desarrollo

$$(\mu_1 - \mu_2)^\tau \mathbf{\Sigma}_{XX}^{-1}(\mu_1 + \mu_2) = \mu_1^\tau \mathbf{\Sigma}_{XX}^{-1}\mu_1 - \mu_2^\tau \mathbf{\Sigma}_{XX}^{-1}\mu_2 \tag{2.18}$$

Que sigue

$$L(\mathbf{x}) = \log_e\left(\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})}\right) = b_0 + \mathbf{b}^\tau\mathbf{x} \tag{2.19}$$

El cual representa una función lineal de $\mathbf{x}$ donde:

$$\mathbf{b} = \mathbf{\Sigma}_{XX}^{-1}(\mu_1 + \mu_2) \tag{2.20}$$

$$\mathbf{b}_0 = -\frac{1}{2}\left(\mu_1^\tau \mathbf{\Sigma}_{XX}^{-1}\mu_1 - \mu_2^\tau \mathbf{\Sigma}_{XX}^{-1}\mu_2\right) + \log_e\left(\frac{\pi_1}{\pi_2}\right) \tag{2.21}$$

Así, si

$$L(\mathbf{x}) > 0 \tag{2.22}$$

La observación $\mathbf{x}$ se asigna a la clase Π_1, de lo contrario a Π_2.

La regla definida en 2.22 es conocida como *análisis discriminante lineal gausiano* (Gaussian Linear Discriminant Analysis) o simplemente LDA.

2.4.4. Regresión logística (Logistic Regression, LR)

Se usa el termino regresión logística, debido a que se trata de un modelo de regresión generalizada, a veces también se le denomina discriminación logística, porque en realidad este método sirve para resolver problemas de clasificación [29].

La regresión logística es un tipo especial de regresión que se utiliza para explicar y predecir una variable categórica binaria (dos grupos) en función de varias variables independientes que a su vez pueden ser cuantitativas o cualitativas. Permite modelar la probabilidad de que ocurra un evento dado una serie de variables independientes.

Sea y_i una variable con respuesta binaria, y p la probabilidad de clase:

$$p = \begin{cases} p_i, & \text{si } y_i = 1 \\ 1 - p_i, & \text{si } y_i = 0 \end{cases} \tag{2.23}$$

La variable respuesta sigue una distribución binomial de parámetros 1 y p_i, donde... , La parte sistemática del modelo es esta probabilidad:

$$\mu_i = E[y_i] = 1 * p_i + 0 * (1 - p_i) = p_i \tag{2.24}$$

De donde se tiene que la función de unión r^{-1} debe realizar una transformación del predictor lineal en el intervalo $[0, 1]$. Para ello la función más utilizada es la *función logística*:

$$p_i = \frac{1}{1 + e^{-\beta' x_i}} \tag{2.25}$$

Y esta relación también puede ser escrita como:

$$\log\left(\frac{p_i}{1 - p_i}\right) = \beta' x_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} \tag{2.26}$$

Y este es el predictor lineal clásico.

Para estimar los coeficientes del modelo, de busca maximizar la función de verosimilitud:

$$l(b) = \sum_{i=1}^{n} \log p_i = \sum_{i=1}^{n} \left(y_i \log p_i + (1 - y_i) \log(1 - y_i)\right) \tag{2.27}$$

$$l(b) = \sum_{i=1}^{n} \left(y_i \log\left(\frac{p_i}{1 - p_i}\right) + \log(1 - y_i)\right) \tag{2.28}$$

Y sustituyendo el resultado de la función de unión queda:

$$l(b) = \sum_{i=1}^{n} \left(y_i b' x_i - \log(1 + e^{b' x_i})\right) \tag{2.29}$$

Obteniendo las derivadas parciales con respecto a b:

$$\frac{\partial l(b)}{\partial b} = \sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} \left(\frac{1}{1 + e^{-b' x_i}}\right) x_i \tag{2.30}$$

Que en notación matricial sería:

$$\frac{\partial l(b)}{\partial b} = \mathbf{X}'(\mathbf{y} - \mathbf{p}) \tag{2.31}$$

Y obteniendo las segunda derivada se llega a

$$\frac{\partial^2 l(b)}{\partial b \partial b'} = \sum_{i=1}^{n} \left(\frac{e^{-b' x_i}}{(1 + e^{-b' x_i})^2}\right) x_i x_i' \tag{2.32}$$

$$\frac{\partial^2 l(b)}{\partial b \partial b'} = \sum_{i=1}^{n} \left(p_i(1 - p_i)\right) x_i x_i' \tag{2.33}$$

Que en notación matricial sería:

$$\frac{\partial^2 l(b)}{\partial b \partial b'} = -\mathbf{X}'\mathbf{W}\mathbf{X} \tag{2.34}$$

donde $\mathbf{W}$ se define como:

$$\mathbf{W} = \begin{bmatrix} \ddots & & \\ & p_i(1 - p_i) & \\ & & \ddots \end{bmatrix} \tag{2.35}$$

Y con una iteración de Newton-Raphson se tiene:

$$\mathbf{b}^{t+1} = \mathbf{b}^t + (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'(\mathbf{y} - \mathbf{p}) = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{z} \tag{2.36}$$

Con $\mathbf{z} = \mathbf{X}\mathbf{b}^t + \mathbf{W}^{-1}(\mathbf{y} - \mathbf{p})$, de donde es posible obtener $\mathbf{b}$.

Y este proceso es iterativo hasta al convergencia del algoritmo. Al final el vector p contiene las probabilidades por medio de las cuales se predice la clase de un nuevo caso.

2.5. Rendimiento de Clasificadores

En esta sección se presentan algunas prácticas utilizadas para la evaluación del rendimiento de algoritmos de clasificación binaria.

2.5.1. Validación Cruzada

En muchas ocasiones los ejemplos a los que se puede tener acceso para llevar a cabo una investigación constituyen un conjunto con cardinalidad. En el caso de aprendizaje automático, para evaluar los resultados de un estudio, es necesario comprobar que éstos son independientes de la partición entre datos de entrenamiento y prueba.

Un técnica conocida y comúnmente utilizada en aprendizaje automático para la verificación de resultados es la *validación cruzada* de k conjuntos. Esta técnica consiste en repetir y calcular la media aritmética de las métricas de evaluación sobre diferentes particiones. Se puede decir que la validación cruzada consiste en realizar k diferentes experimentos, obteniendo las estadísticas de cada uno y entregado métricas de rendimiento conjunta de todos los experimentos.

Una forma de realizar la validación cruzada es mediante la estrategia conocida como k **conjuntos (k-fold)**. Esta estrategia sugiere dividir el conjunto de datos de tamaño N en k subconjuntos $T_1, T_2, \ldots, T_k$ de igual tamaño y seleccionar uno por uno cada subconjunto para realizar la validación. El resto de los $k-1$ subconjuntos son destinados para entrenamiento. De esta forma cada subconjunto es utilizado $k-1$ veces en la etapa de entrenamiento y 1 vez para validación del modelo.

La figura 2.8 bosqueja el funcionamiento de la validación cruzada. En ella se puede observar como un subconjunto del total de características extraídas es utilizado $k-1$ veces para entrenar y 1 para pruebas.

Para cada uno de los experimentos E_k se obtienen las métricas de desempeño D_k y se calcula un único conjunto de medidas D_T de acuerdo a la siguiente ecuación:

$$D_T = \frac{1}{k} \sum_{i=1}^{k} D_i \tag{2.37}$$

Que como puede observarse, no es más que la media aritmética de las medidas de desempeño de cada experimento individual.

Si para la validación cruzada se define un total de conjuntos k igual al número de ejemplos disponibles N se dice que se está realizando una validación cruzada **dejando uno fuera (leave-one-out)**. Este caso particular de la validación de k-conjuntos, tiene la variante que se utiliza un único elemento para validar el proceso completo, lo que se

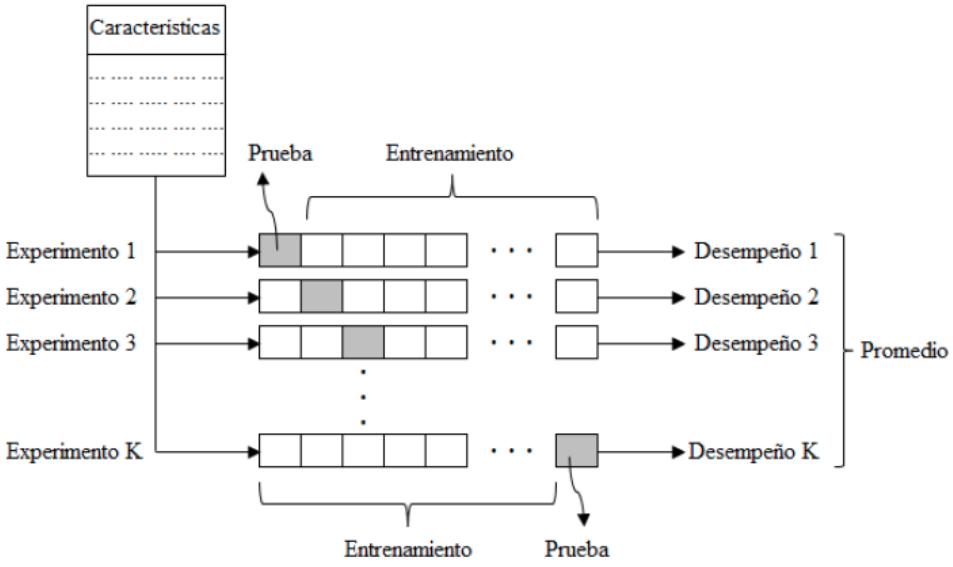


Figura 2.8: Validación cruzada de k conjuntos

refleja en métricas de desempeño de todo o nada. Este proceso se repite un número de veces que es igual al número de ejemplos con los que se cuenta, utilizando exactamente una vez cada patrón para la validación.

2.5.2. Matriz de confusión

La *matriz de confusión (contingency table)* es una forma de evaluar los resultados generados por un clasificador. Su nombre proviene del hecho que su contenido es muestra de si el clasificador está confundiendo las clases para los datos suministrados como entrada. En un caso como el de lesiones de cáncer de mamá, donde se desea saber si una lesión es maligna o no, es decir, si la respuesta del clasificador es positiva (P) o negativa (N). Los cuatro estados que sirven de base para el análisis de rendimiento a continuación son:

- **Verdadero positivo (VP):** La patología fue clasificada como positiva y es positiva.
- **Falso positivo (FP):** La patología fue clasificada como positiva y es negativa.
- **Falso negativo (FN):** La patología fue clasificada como negativa y es positiva .
- **Verdadero negativo (VN):** La patología fue clasificada como negativa y es negativa.

Y normalmente se organizan como se muestra en la tabla 2.7:

	Real Positivo (PP)	**Real Negativo (PN)**
Predicción Positivo (RP)	Verdadero positivo (VP)	Falso positivo (FP)
Predicción Negativo (RN)	Falso negativo (FN)	Verdadero negativo (VN)

Cuadro 2.7: Matriz de confusión

Las métricas básicas utilizadas para medir el desempeño de un clasificador son: *éxito (Accuracy)* o porcentaje de éxito (ver ecuación 2.38), *error* (ver ecuación 2.39), *sensibilidad (Sensitivity, Recall)* o tasa de verdaderos positivos (TVP) (ver ecuación 2.40) y *especificidad (Specificity, Inverse Recall)* o tasa de verdaderos negativos (TVN) (ver ecuación 2.41). Éstas cuatro medidas de desempeño se calcular a partir de la información proporcionada en la matriz de confusión, y tienen un rango de valores entre 0 y 1.

$$éxito = \frac{VP + VN}{VP + VN + FP + FN} \tag{2.38}$$

$$error = \frac{FP + FN}{VP + VN + FP + FN} \tag{2.39}$$

$$sensibilidad = \frac{VP}{VP + FN} \tag{2.40}$$

$$especificidad = \frac{VN}{VN + FP} \tag{2.41}$$

Al encontrarse en un rango de entre 0 y 1, la sensibilidad y especificidad pueden ser interpretadas como la probabilidad que un ejemplo *predicho positivo* esté correctamente clasificado y que uno *predicho negativo* sea realmente un caso negativo, respectivamente. De lo anterior se puede resaltar que el clasificador ideal tendría una sensibilidad y especificidad de 1.

Otra observación que se debe resaltar es que la matriz de confusión tiene 3 grados de libertad, es decir, una vez determinados 3 de los 4 valores que forman la matriz, el cuarto y último puede ser obtenido a partir de los otros. Esto último puede resultar de utilidad al interpretar los valores de las métricas de desempeño descritas en párrafos anteriores.

2.5.3. Curva ROC (Receiver Operating Characteristics)

Un análisis importante que permite realizar una comparación entre diferentes algoritmos es el denominado *Curva ROC (Receiver Operating Characteristics, ROC)*. En su forma más simple, la curva ROC es gráfica en un espacio bidimensional definido por la *tasa de verdaderos positivos* como función de la*tasa de falsos positivos*.

Un clasificador es preferible con respecto a otro si su curva ROC es muy cercana a

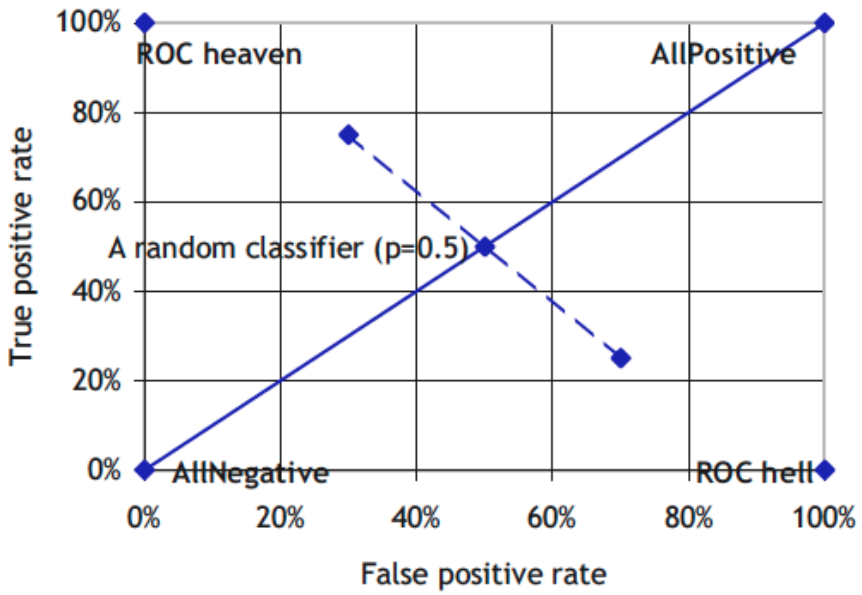


Figura 2.9: Espacio ROC

su umbral de rendimiento. Una de las ventajas de este tipo de medida de rendimiento radica en el hecho que permite la comparación visual entre algoritmos al estar basada en una representación netamente geométrica.

2.6. Bases de datos

En esta sección se presentan 3 bases de datos que han sido utilizadas en investigaciones afines a la presente. Estas bases son: Mammographic Image Analysis Society (MIAS), Digital Database for Screening Mammography (DDSM) y Image Retrieval in Medical Applications(IRMA).

2.6.1. Mammographic Image Analysis Society (MIAS)

La base de datos de mamografías que se ha utilizado es la proporcionada por la Mammographic Image Analysis Society (MIAS) en su versión de libre acceso denominada miniMIAS. La base de datos miniMIAS contiene 322 imágenes de mamografías digitales, numeradas de menor a mayor donde cada par de mamografías representa a la mama izquierda (número de archivo par) y a la mama derecha (número de archivo impar) de una sola paciente. Las imágenes fueron digitalizadas a 200 μ /pixel y el tamaño de cada archivo es de 1024 pixeles x 1024 pixeles. De las 322 imágenes 7 presentan dos anomalías por imagen, 1 presenta tres anomalías por imagen y las restantes presentan solo una anomalía, dando un total de 330 hallazgos [56].

2.6.2. Digital Database for Screening Mammography (DDSM)

La base de datos Digital Database for Screening Mammography (DDSM) es el resultado de la colaboración realizada por Massachusetts General Hospital, Sandia National Laboratories y el Departamento de Ciencias Computacionales e Ingeniería de la Universidad del Sur de la Florida. La base de datos fue completada en el otoño del año 1999 y desde entonces ha sido extensamente utilizada en estudios diversos como [26, 27]. La DDSM está organizada en casos y volúmenes. Un volumen es una colección de casos, los cuales han sido agrupados para facilitar la distribución de los mismos. La base de datos está constituida por 2632 casos de estudio, cada uno de los cuales contiene dos imágenes de cada mama, que corresponden a las vistas craniocaudal (CC) y mediolateral oblicua (MLO) de cada paciente. La DDSM agrupa un total de 10528 mamografías, que fueron digitalizadas a partir de las placas radiográficas originales mediante el uso de cuatro escáneres diferentes que se describen en la tabla 2.8.

Escáner	Muestreo (microns)	Nivel de gris (bits)
DBA M2100 ImageClear	42	16
Howtek 960	43.5	12
Lumisys 200 Laser	50	12
Howtek MultiRad850	43.5	12

Cuadro 2.8: Escáneres utilizados para la construcción de la DDSM

Todos los casos de estudio agrupados en DDSM fueron efectuados entre octubre de 1988 y febrero de 1999.

Casos de estudio

Un caso de estudio en la DDSM contiene entre 6 y 10 archivos relacionados, almacenados en un directorio separado para cada caso. Los directorios se dividen en 3 categorías:

- **Normal:** 702 casos de mamografías sin lesiones.
- **Benign:** 1011 casos con lesiones benignas, dentro de los cuales 141 contienen anotaciones.
- **Cancer:** 919 casos de lesiones diagnosticadas con cáncer.

En la figura 2.10 se presenta la información sobre los archivos contenidos en el directorio del caso 3024.

La información contenida dentro de cada directorio se desglosa a continuación:

- 1 archivo ".ics": Texto pleno que contiene información de la paciente y de las imágenes de cada caso.

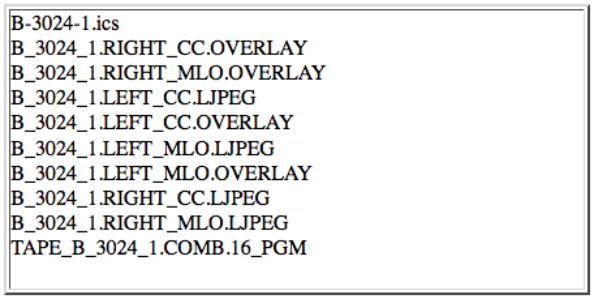


Figura 2.10: Lista del diretorio del caso 3024

- 4 archivos ".LJPEG": Imágenes comprimidas que representan las vistas CC y MLO de mama izquierda y derecha.
- 0-4 archivos ".OVERLAY": Archivos que describen las lesiones encontradas por un radiólogo experto en las mamografías de cada caso.
- 1 archivo ".16_PGM": Versión que concatena en una sola imagen 16 bits/pixel con formato ".pgm" las vistas CC y MLO de mama izquierda y derecha.

Estos archivos contienen información sobre el tipo de digitalizador utilizado, así como del estado de la lesión detectada en el seno de acuerdo a los establecido por el *American College of Radiology* (ACR).

La tabla 2.9 muestra el detalle de todos los volúmenes que componen la DDSM.

Archivo "ics"

Es un archivo de texto en formato ASCII, en el cual se lista la información relacionada al caso en cuestion como la fecha en la que se llevo a cabo el estudio, edad del paciente en el momento del estudio, fecha de digitilización del caso, el digitalizador que se utilizó y la lista completa de los archivos relacionados con el caso en concreto.

Dentro del archivo ".ics" se incluye además una calificación, que toma valores entre 1 y 4, para representar la densidad del tejido del seno; ésta calificación fue asignada por un radiólogo experto con base en lo establecido por el ACR.El tamaño de cada archivo de imagen, número de bits por pixel, la resolución de escaneo (en microns) y la información sobre la existencia o falta de archivos tipo ".overlay" vinculados con el caso, completan los datos que se describen en el ".ics". Como puede observarse en la figura 2.11, cada una de las cuatro imágenes que componen el caso 3024 disponen de archivos ".overlay". Éstos archivos se listan en la figura 2.10. Si en las líneas de descripción de la imagen se especifica "NON-OVERLAY" en lugar de "OVERLAY", esa imagen específica del caso en cuestión no cuenta con archivo ".overlay" asociado.

Casos	Nombre del volumen
118 casos	normal_01
117 casos	normal_02
38 casos	normal_03
57 casos	normal_04
47 casos	normal_05
60 casos	normal_06
78 casos	normal_07
27 casos	normal_08
59 casos	normal_09
23 casos	normal_10
58 casos	normal_11
20 casos	normal_12
69 casos	cancer_01
88 casos	cancer_02
66 casos	cancer_03
31 casos	cancer_04
83 casos	cancer_05
56 casos	cancer_06
52 casos	cancer_07
60 casos	cancer_08
81 casos	cancer_09
59 casos	cancer_10
59 casos	cancer_11
80 casos	cancer_12
21 casos	cancer_13
42 casos	cancer_14
72 casos	cancer_15
80 casos	benign_01
69 casos	benign_02
64 casos	benign_03
81 casos	benign_04
62 casos	benign_05
74 casos	benign_06
61 casos	benign_07
64 casos	benign_08
75 casos	benign_09
21 casos	benign_10
62 casos	benign_11
64 casos	benign_12
72 casos	benign_13
21 casos	benign_14
75 casos	benign_without_callback_01
66 casos	benign_without_callback_02

Cuadro 2.9: Resumen de volumenes que contiene la DDSM

```
ics_version 1.0
filename B-3024-1
DATE_OF_STUDY 2 7 1995
PATIENT_AGE 42
FILM
FILM_TYPE REGULAR
DENSITY 4
DATE_DIGITIZED 7 22 1997
DIGITIZER LUMISYS
SELECTED
LEFT_CC LINES 4696 PIXELS_PER_LINE 3024 BITS_PER_PIXEL 12 RESOLUTION 50 OVERLAY
LEFT_MLO LINES 4688 PIXELS_PER_LINE 3048 BITS_PER_PIXEL 12 RESOLUTION 50 OVERLAY
RIGHT_CC LINES 4624 PIXELS_PER_LINE 3056 BITS_PER_PIXEL 12 RESOLUTION 50 OVERLAY
RIGHT_MLO LINES 4664 PIXELS_PER_LINE 3120 BITS_PER_PIXEL 12 RESOLUTION 50 OVERLAY
```

Figura 2.11: Contenido del archivo B-3024-1.ics

Archivos "lossless JPEG"

Las imágenes contenidas en DDSM están codificadas y comprimidas siguiendo el estándar Lossless Joint Pictures Expert Group (LJPEG). Aún después de comprimidas las imágenes de la DDSM continúan siendo archivos de gran tamaño, esto debido a la resolución de los escáneres utilizados en el proceso de digitalización (tabla 2.8). Los encargados de la base de datos proveen un software especial, escrito en lenguaje C y que corre en la plataforma SunOS 5.5, para descomprimir las imágenes.

Archivos "overlay"

Los casos con anomalías, es decir, aquellas mamografías donde se ha identificado algún tipo de lesión cuentan con un archivo de tipo ".overlay", el cual contiene información de cada lesión detectada. Como cada caso incluye cuatro mamografías que son el resultado de la combinación entre vistas y mama, un caso puede tener desde cero archivo ".overlay" hasta un máximo de cuatro. Para determinar si una imagen de un caso en concreto cuenta con archivo ".overlay" basta con revisar el ".ics" del caso, en el cual, en las líneas correspondientes a la información de las imágenes se incluye la leyenda "OVERLAY" o "NON-OVERLAY", si existe o no un archivo de este tipo asociado a esa imagen. La figura 2.12 muestra el contenido del archivo ".overlay" del caso 3024 para la vista CC del seno derecho.

En un archivo ".overlay" se puede especificar una o más lesiones. En la primera línea del archivo se indica el total de lesiones detectadas en la imagen en cuestión. En el caso que una imagen congregue más de una lesión (TOTAL_ABNORMALITIES mayor a 1), la información de cada una de ellas es listada una después de la otra dentro del archivo.

Para cada lesión detectada en una mamografía se especifica el tipo de lesión (LESION_TYPE), el nivel de apreciación (ASSESSMENT), la sutileza de la lesión (SUBLETY), la patología (PATHOLOGY) y el número de contornos (TOTAL_OUTLINES). El tipo de lesión, apreciación, sutileza y patología de los casos contenidos en la DDSM fueron determinados por un radiólogo experto. De la misma forma, los contornos fueron

```
TOTAL_ABNORMALITIES 1
ABNORMALITY 1
LESION_TYPE CALCIFICATION TYPE PLEOMORPHIC-FINE_LINEAR_BRANCHING DISTRIBUTION REGIONAL
ASSESSMENT 5
SUBTLETY 4
PATHOLOGY MALIGNANT
TOTAL_OUTLINES 4
BOUNDARY
8 1368 4 4 4 4 4 4 4 4 2 2 2 2 2 2 2 2  ...  0 0 0 0 0 0 0 0 0 1 #
        CORE
168 1824 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2  ...  1 0 1 1 0 1 1 0 1 1 #
        CORE
384 1848 2 2 2 2 2 2 2 2 1 1 1 1 1 1 1 1  ...  0 0 0 0 0 0 0 0 0 0 #
        CORE
368 2192 6 6 6 6 6 6 6 6 0 0 0 0 0 0 0 0  ...  0 0 0 0 0 0 0 0 0 0 #
```

Figura 2.12: Contenido del archivo B_3024_1.RIGHT_CC.OVERLAY

trazados por el medico especialista. En los casos donde se presentan más de un contorno, el primer contorno es especificado por la palabra "OUTLINE" y los siguiente por la palabra "CORE", siendo el primer contorno el que contiene al resto. En todos los casos, un contorno es especificado mediante una cadena de caracteres que inicia después de una alguna de las palabras "OUTLINE" o "CORE", según sea el caso y termina con el símbolo #.

El contorno completo de la lesión es especificado en forma de cadena de caracteres. Para interpretar la cadena que representa a un contorno es necesario aclarar que el pixel origen de la imagen se ubica en la esquina superior izquierda. Los primeros dos valores de la cadena representan las coordenadas del punto de inicio del contorno de la lesión. Los siguientes caracteres indican un desplazamiento hacía alguno de los píxeles vecinos del pixel actual, como se observa en la figura 2.13(a). El moverse hacía alguno de los píxeles vecinos implica modificar las coordenadas del píxel actual tal como se muestra en la figura 2.13(b).

		X	
	7	0	1
Y	6	**P**	2
	5	4	3

(a) Píxeles vecinos

	0	1	2	3	4	5	6	7
X	0	1	1	1	0	-1	-1	-1
Y	-1	-1	0	1	1	1	0	-1

(b) Movimientos absolutos

Figura 2.13: Interpretación de la cadena de caracteres del contorno de una lesión.

Archivo "16-bit PGM"

En el archivo ".16_PGM" se concatenan las vistas CC y MLO de mama izquierda y derecha en una sola imagen con formato 16-bit PGM (Portable Gray Map). La calidad

de las imágenes concatenadas en este archivo es baja, pero el objetivo de dicho archivo es proveer una vista rápida de las mamografías asociadas al caso en cuestión.

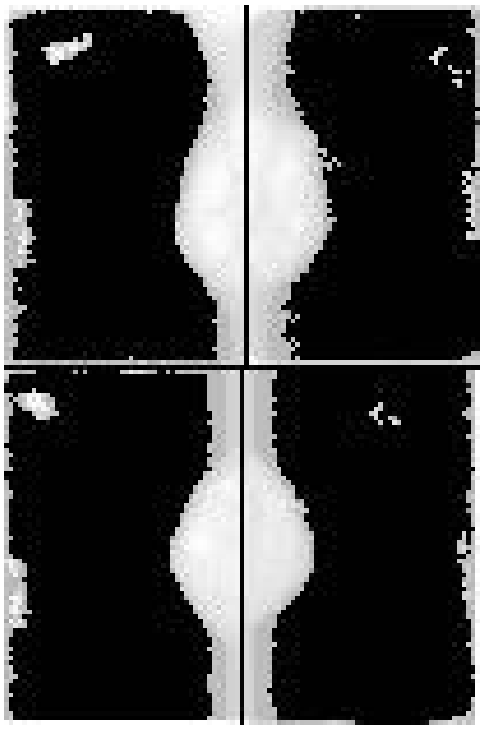


Figura 2.14: Imagen TAPE_B_3024_1.COMB.16_PGM que resume el caso 3024

Demografía de los casos

La mayoría de los casos contenidos en la DDSM proviene del programa de mamografías del *Massachusetts General Hospital* ubicado Boston, Estados Unidos. La otra gran parte de las imágenes en la base de datos tienen origen en el programa de mamografías de la *Wake Forest University School of Medicine* localizado en Winston-Salem, Estados Unidos. La totalidad de los casos en la DDSM provienen de pacientes mujeres. La tabla 2.10 desglosa los pacientes por raza y hospital donde fue efectuado el estudio:

	MGH	**WFUSM**
Asiatica	2.06	0.2
Hispánica	6.55	1.8
Nativa americana	0.00	0.1
Africana	4.12	20.4
Caucasia	56.18	77.0
Otra	0.75	0.1
Desconocida	30.34	0.3

Cuadro 2.10: Desglose de casos por raza y hospital de estudio

2.6.3. Image Retrieval in Medical Applications(IRMA)

El proyecto Image Retrieval in Medical Applications (IRMA) es una colaboración entre Departamento de Radiología, de la División para el Procesamiento de Imágenes Médicas del Departamento de Informática Médica y la Cátedra de Ciencias Computacionales VI de la Aachen University of Technology (RWTH Aachen), ubicada en Alemania. El objetivo de dicho proyecto es el desarrollo y la implementación de métodos de alto nivel para la recuperación de imágenes basadas en su contenido (content-based image retrieval) con aplicaciones, a nivel prototipo, para el diagnóstico medico usando imágenes de radiografías.

La base de datos del proyecto IRMA cuenta actualmente con más de 30000 imágenes medicas con información sobre el contenido visual de las mismas. En el caso de las mamografías, la base congrega más de 10000 casos, que son el resultado de la unión de las bases The Mammographic Image Analysis Society Digital Mammogram Database (MIAS), The Digital Database for Screening Mammography (DDSM), The Lawrence Livermore National Laboratory (LLNL) adicionado con algunos casos provenientes de la propia institución, Rheinische-Westfälische Technische Hochschule (RWTH) Aachen.

En el sitio de Internet, el proyecto IRMA ofrece una versión para descargar de la base Digital Database for Screening Mammography (DDSM), de la Univerdidad de la Florida [26]. Las imágenes que IRMA pone a disposición se encuentran previamente convertidas a formato PNG de 16bit y están almacenadas en archivos TAR y comprimidas con GNU Zip. La organización de las imágenes mantiene el orden de la de la base DDSM, lo cual se describe en la sección 2.6.2.

2.7. Trabajos Relacionados

En esta sección se detallan algunos trabajos que comparten el mismo problema pero atacado con una estrategia diferente o bien que buscan resolver problemas afines con técnicas relacionadas a las que se incluyen en el modelo de solución de la presente propuesta.

Tang et al. Realizan una descripción de los avances recientes en el desarrollo de sistemas CAD. Entre los elementos importantes presentados por los autores se encuentra la detección y segmentación de señales en las imagenes, un proceso de extracción de caracteristicas y la clasificación de los mismos. Los autores exponen que estos procesos o etapas constituyen la base de lossistemas que en la actualidad ya se comercializan, así como la mayoria sistemas que se desarrollan en investigación. [57].

Oporto Díaz realizó como tesis de doctorado, bajo la asesoría del Dr. Hugo Terashima, un estudio sobre detección automática de agrupamientos de microcalificaciones en mamografías digitalizadas [45]. El autor realizó el pre-procesamiento de mamografías utilizando un filtro Gaussiano, identificando específicamente microcalcificaciones y grupos de microcalcificaciones, las que fueron posteriormente clasificadas utilizando redes

neuronales.

El proceso de selección de características constituye una etapa importante del diseño de clasificadores de lesiones de cáncer de mama con base a mamografías digitalizadas. El objetivo de esta reducción de dimensiones tiene fin encontrar el subconjunto de características que aportan mayor poder de discriminación para decir si una lesión es maligna o no. En [66] los autores presentan un método denominado PSO-kNN, el cual esta basado en optimización la optimización de enjambre de particulas (Particle Swarm Optimization, PSO), que permite seleccionar los parámetros de forma heurística utilizando la técnica de k vecinos más cercanos (k-nearest neighbor, k-NN) para clasificar lesiones.

Una vez determinado el conjunto de características a utilizar, se procede a la etapa de clasificación de lesiones en malignas o benignas. Varias técnicas han sido exploradas con este fin, entre las que se pueden resaltar las máquinas de vectos de soporte (SVM), redes neuronales artificiales (NN), redes neuronales artificales evolutivas (EANN) entre otras. Hernández et al. [15] presentan un panorama del uso de las estrategias de computación evolutiva en aplicaciones médicas, específicamente en la selección de características para la clasificación de microcalcificaciones en mamografías digitales, utilizando técnicas como algoritmo genéticos y búsqueda secuencial. De la misma manera, el estudio presentado en [16], aborda el problema de selección de características utilizando un modelo basado en un algoritmo genético (GA) para seleccionar las características con mayor poder de discriminación de lesiones de microcalcificaciones y agrupaciones de microcalcificaciones.

En [10] se presenta el desarrollo de un modelo de detección, segmentación y clasificación de masas tumorales en mamografías digitales. El estudio desarrollado por Castro Astudillo, realizado sobre imágenes de la base *miniMIAS* [56], inicia con la detección de lesiones detección de bordes de una señal utilizando transformadas Wavelet y posteriormente mediante algoritmos de región creciente se refina la línea detectada como borde. A cada una de éstas señales, que representan una posible masa, se le extraen 50 características, con las que se decide de forma automática si una señal representa o no una masa. De la misma manera, se extrae el mismo conjunto de 50 características para decidir si una masa tumoral es maligna o no y posteriormente se lleva a cabo un proceso de reducción de dimensiones utilizando un enfoque de *envoltura* donde la búsqueda se lleva a cabo mediante un algoritmo genético y la clasificación se realiza con una red neuronal. Se reportan estadísticas de rendimiento para la clasificación de de señales en masas y sobre la malignidad de las masas en los casos de la base miniMIAS.

Campanini et al. Describen en [9] el uso de SVM para diagnóstico asistido por computadora. Ellos comentan que los clasificadores comúnmente utilizados para sistemas CAD incluyen redes neuronales y arboles de decisiones, y resaltan las limitantes de estos dos algoritmos de clasificació para atacar problemas no lineales, principalmente cuando se cuenta con un alto nivel de dimensionalidad en los patrones. Este trabajo incluye referencias donde se ha utilizado SVM tanto para detección como para clasificación de

lesiones de cáncer de mama.

Un estudio detallado sobre el rendimiento de diferentes algoritmos de clasificación es presentado por Kumar en On the Classification of Imbalanced Datasets [38] donde se incluyen técnicas como SVM, ANN, k-NN y clasificación bayesiana.

Bourdes et al en [5] realizan una comparación entre los algoritmos de redes neuronales y regresión logística para predecir cáncer de mama. El estudio resalta la selección de la regresión logística como una medida estándar de predicción en bioestadística y justifica la comparación de rendimiento de este algoritmo con las redes neuronales. De manera similar Chhatwal et al [14] presentan un una investigación donde utilizan dos modelos de regresión logística para predecir el riesgo de cáncer de mama con base en la información de la *National Mammography Database*.

El cuadro 2.11 muestra una lista de los trabajos de investigaciones relacionadas con la presente tesis.

Autor	**Título**	**Año**
Kumar, A. et al [38]	On the Classification of Imbalanced Datasets	2012
Castro, C. et al [10]	Detección, etiquetado y reconstrucción de masas y su clasificación por medio de redes neuronales en mamografías digitales	2012
Zyout, I. et al [66]	Embedded feature selection using pso-knn: Shape-based diagnosis of microcalcification clusters in mammography.	2011
Hernández, R. et al [16]	Feature Selection for the Classification of Digital Mammograms using Genetic Algorithms, Sequential Search and Class Separability.	2010
Verma, B. et al [63]	Classification of benign and malignant patterns in digital mammograms for the diagnosis of breast cancer	2010
Zhang, Y. et al [64]	Using BI-RADS Descriptors and Ensemble Learning for Classifying Masses in Mammograms	2010
Chhatwal, J. et al [14]	A Logistic Regression Model Based on the National Mammography Database Format to Aid Breast Cancer Diagnosis	2009
Tang, J. et al [57]	Computer-aided detection and diagnosis of breast cancer with mammography: recent advances	2009
Bourdes, V. et al [5]	Breast Cancer Predictions by Neural Networks Analysis: a Comparison with Logistic Regression	2007
Hernández, R. et al [28]	Evolutionary Neural Networks Applied To The Classification Of Microcalcification Clusters In Digital Mammograms	2006
Campanini, R. et al [9]	Support Vector Machines in CAD Mammography	2006
Oporto, S. et al [45]	Detección automática de agrupamientos de microcalificaciones en mamografías digitalizadas	2004

Cuadro 2.11: Trabajos relacionados

2.8. Resumen

Las causas que originan el cáncer de mama son muy diversas. La severidad de la enfermedad se expresa en etapas, que va desde la etapa 0 o etapa inicial, hasta la etapa IV que representa mayor gravedad. La mamografía es considerada como una técnica de detección efectiva y es la más usada por las ventajas que brinda. Uno de los hallazgos que se puede encontrar en una mamografía son las masas tumorales benignas o malignas y se describen por su forma, márgenes que la rodea y su densidad. Cuanto más irregular la forma y márgenes de una masa, mayor probabilidad que se trate de cáncer.

Los sistemas de diagnóstico asistido por computadora (CAD) reciben como entrada una imagen médica, como la mamografía y entregan como resultado un diagnóstico o predicción sobre el contenido de la imagen en cuestión. El proceso de un sistema CAD inicia con el pre-procesamiento y una técnica para ellos se basa en la uso de un filtro de mediana para eliminar ruido contenido en la imagen. La etapa de segmentación entrega como resultado la región de interés que contiene la lesión encontrada en la mamografía. Esta información visual puede ser convertida a numérica mediante la extracción de características. Mediante la información que representan las características de una lesión se puede emitir de manera automática un diagnóstico sobre la malignidad de la lesión.

Existen algoritmos computacionales capaces de aprender de ejemplos y que pueden ser utilizados para realizar labores de predicción automática. Estos algoritmos reciben un conjunto de entradas que representan la información pertinente al dominio en cuestión. Existen diferentes procesos que se utilizan para refinar el conjunto de entradas que se le suministran a un algoritmo clasificador, esto se traduce en beneficios en tiempo de cómputo y en resultados más precisos.

Capítulo 3

Modelo de Solución

El objetivo de la presente investigación es desarrollar un procedimiento que apoye en el diagnóstico de cáncer de mama con la ayuda de una herramienta de clasificación de lesiones sobre mamografías digitales. Para alcanzar dicho objetivo, se ha definido un modelo de solución que recibe como entrada una imagen digital, en este caso una mamografía, y la información relacionada con la lesión detectada en dicha imagen, y posteriormente, mediante la aplicación de algoritmos computacionales se obtiene un conjunto de características que se utilizan para generar de manera automática un diagnostico de cada lesión contenida en la imagen. El presente capítulo detalla el modelo de solución propuesto y adicionalmente describe los algoritmos que se desarrollaron en cada una de las fases que lo conforman.

La figura 3.1 presenta de forma gráfica el modelo de solución propuesto para este trabajo de investigación. Inicialmente se obtienen las regiones de interés mediante la segmentación de las lesiones contenidas en las imágenes de los casos de estudio obtenidos de la base de datos Digital Database for Screening Mammography DDSM. Para todas las lesiones de los casos de estudio seleccionadas se calculan los valores de un conjunto de características que representan de manera cuantitativa la información visual contenida en la imagen y relevante en cuanto a lesión. La siguiente etapa del modelo de solución consiste de reducir el conjunto de características extraídas hasta el subconjunto de menor tamaño que aporte el mayor poder para discriminar entre lesiones malignas y benignas. La etapa de clasificación detalla los aspectos de los diferentes algoritmo de aprendizaje automático que se consideraron en el modelo propuesto.

3.1. Segmentación de lesiones

El proceso de segmentación extrae una ventana dentro de una imagen, la cual representa la *región de interés (ROI)* que contiene a la lesión detectada en esa imagen en específico. La figura 3.2 bosqueja el proceso de segmentación utilizado en esta investigación. El proceso de segmentación de lesiones inicia con el preprocesamiento de

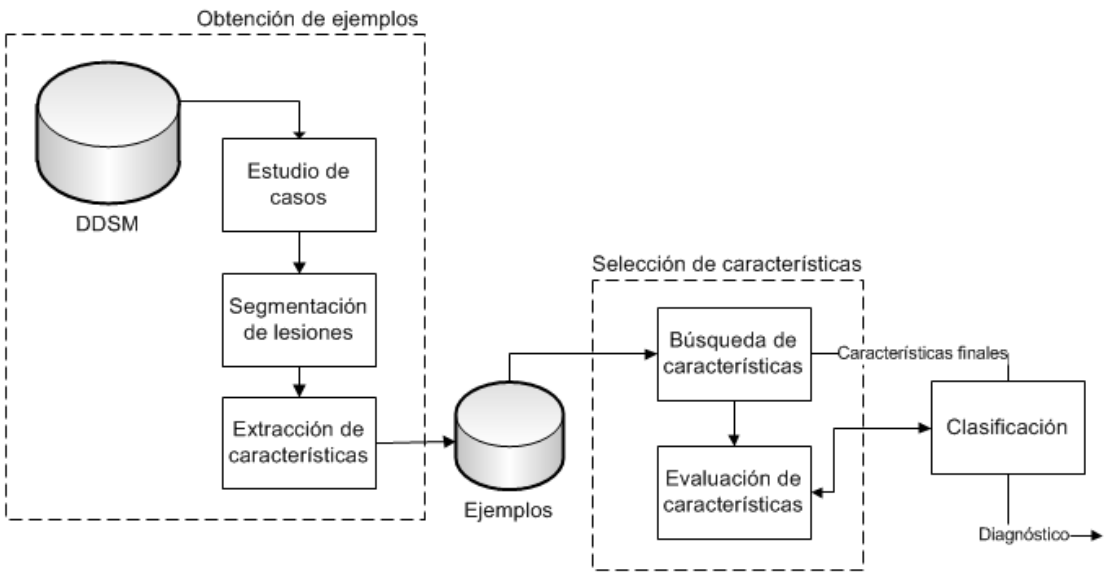


Figura 3.1: Modelo de Solución.

la imagen en cuestión mediante la aplicación de un *filtro de mediana*. Posteriormente y según la información proporcionada por la base de datos DDSM (fuente de todos los casos de estudio considerados en la presente investigación) se realiza la *construcción del borde* que delimita a la lesión. A partir de este borde se realiza una *binarización* con la que se obtiene una máscara que se utiliza para *segmentar* la imagen y obtener la región de interés que contiene a la lesión. A esta fase ingresa la imagen original y genera como resultado una imagen preprocesada. Los detalles de cada etapa se presentan en las siguientes subsecciones.

3.1.1. Filtro de Mediana

Toda imagen digital puede llegar a contener ruido. El ruido no es más que errores que se generar en el proceso de captura o digitalización de cualquier tipo de imagen. Con el objetivo de mejorar imágenes, el filtro mediana ha sido frecuentemente utilizado para eliminar o reducir ruido aleatorio de alta frecuencia sin difuminar los bordes de los objetos [23, 44], siendo ésta la principal razón por la que se emplea este tipo de filtro para reducir el ruido de una mamografía, sin pérdida de información relevante.

Un filtro de mediana es un filtro no lineal que utiliza una ventana, generalmente cuadrada y reemplaza el valor de cada pixel central por la mediana de sus vecinos dentro de dicha ventana. La ventana determina el tamaño del vecindario a considerar para actualizar el valor de cada pixel de la imagen por la mediana de sus vecinos. La mediana de una serie de números es el valor que ocupa la posición del medio cuando se ordenan los valores de forma creciente (o decreciente) [23]. La figura 3.3 muestra un ejemplo de un filtro de mediana con ventana de tamaño 3x3. Como la ventana alberga

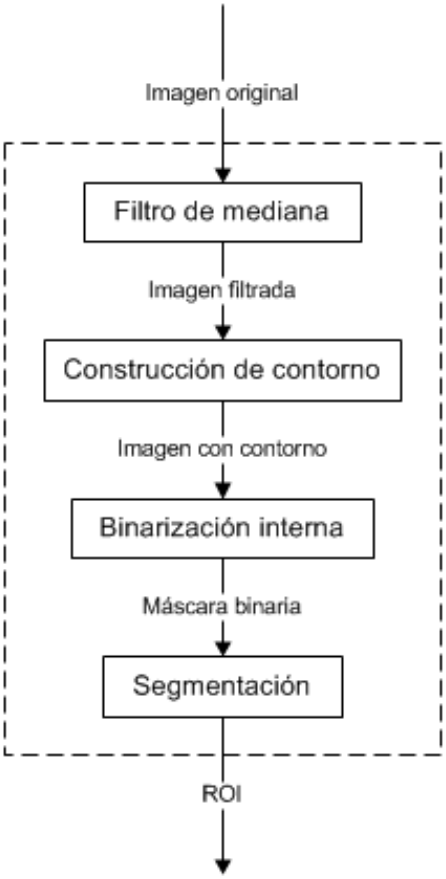


Figura 3.2: Pasos de la fase de extracción de ROI

9 pixeles, el valor del pixel del centro será reemplazado por el valor que se ubique en la quinta posición del orden creciente. Si se tratase de una ventana de tamaño 5x5 el valor del centro de la ventana se reemplazaría por el elemento que ocupe la posición número 13.

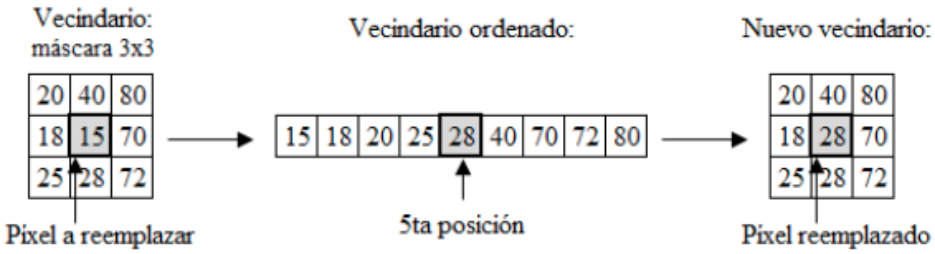


Figura 3.3: Filtro de mediana con ventana de tamaño 3x3

El tamaño de máscara seleccionado y obtenido mediante experimentación en investigaciones previas [10, 45], es de 3x3 ya que reduce el ruido de la mamografía, evitando la pérdida de detalles locales y la alteración del resultado al detectar los bordes de los objetos (ver figura 3.4). Mayor tamaño, como 5x5 o 7x7, puede eliminar las altas frecuencias deshaciéndose así características significativas en la imagen. La figura 3.4(a) muestra la imagen original mientras que la figura 3.4(b) muestra la misma imagen pero con ruido aleatorio. La figura 3.4(c) muestra como el resultado de aplicar el filtro de mediana reduce de mejor manera el ruido a si se aplicara un filtro promedio como en la figura 3.4(d).

(a) Imagen Original (b) Imagen con ruido (c) Filtro de medina (d) Filtro promedio

Figura 3.4: Aplicación del filtro de mediana para eliminar el ruido.

Una vez se ha eliminado de la imagen el ruido que pueda generar errores en los cálculos de las etapas posteriores, se procede a la delimitar la lesión con base en la información proporcionada por la base de datos DDSM.

3.1.2. Construcción de contorno

Por contorno de una lesión se hace referencia al conjunto de pixeles que delimitan la región de la imagen que contiene una lesión. En el caso de la DDSM, el contorno

de las lesiones detectadas en cada caso de estudio se especifica en el respectivo archivo ".overlay" en forma de *cadena de caracteres* (2.6.2). Ésta cadena fue generada a partir de la delimitación de la lesiones por parte del médico especialista. A partir de la información proporcionada por los autores de la DDSM, es posible delimitar dentro de cada caso la lesión previamente detectada.

Los pasos para la construcción del contorno de una lesión a partir de la información disponible en la DDSM se detallan a continuación:

- Lectura de archivo ".overlay".
- Extracción de la cadena de caracteres que representan el contorno de las lesiones.
- Extracción de los primeros dos elementos de la cadena de caracteres, los cuales representan el punto de inicio y fin del contorno.
- Seguir la cadena hasta completar el contorno de la lesión.

3.1.3. Binarización interna

El objetivo de la binarización de la región interna al contorno de la lesión es llegar a tener una imagen binaria que servirá como plantilla para el recorte del área que comprende la lesión para el siguiente apartado.

El método de binarización interna parte de la imagen que contiene el contorno bien delimitado y genera una imagen binaria con valores de 1 en el interior del contorno, incluyendo esta frontera, y valores 0 en la región por fuera del contorno(ver Figura 3.20), para lo que se utiliza el algoritmo de relleno de regiones basado en reconstrucción morfológica [54].

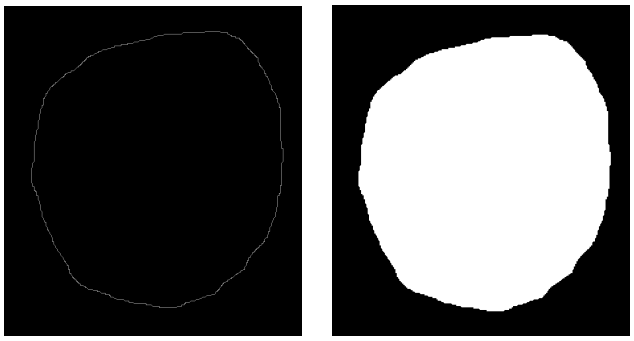
La figura 3.5 muestra el proceso de binarización. La figura 3.5(a) muestra el contorno que delimita la lesión, el cual constituye la entrada de la etapa de binarización interna. La figura 3.5(b) es generada a partir del contorno de la lesión y será utilizada en la siguiente etapa como máscara para extraer la región de interés de la imagen original.

3.1.4. Segmentación

El objetivo de esta etapa del modelo de solución consiste en delimitar la lesión contenida dentro de una imagen, para los pasos posteriores del procesamiento. El proceso de segmentación de lesiones está formado de tres etapas: construcción del contorno de la lesión, binarización interna y segmentación. Estas etapas se describen en las secciones siguientes.

La segmentación se basa en extraer la zona que corresponde a la lesión para luego proceder a realizar cálculos sobre esta área específica.

El método de segmentación parte de la imagen binarizada en el proceso anterior y de la ventana que contiene la lesión. Realiza una copia de los puntos de la imagen solo



(a) Borde de la lesión (b) ROI

Figura 3.5: Proceso de binarización, utilizando la imagen CC del caso 4007 de la base de datos DDSM.

cuando la imagen binarizada tiene un valor de 1, mediante la multiplicación de cada uno de los valores de la imagen que contiene la masa con los valores de la binarizada.

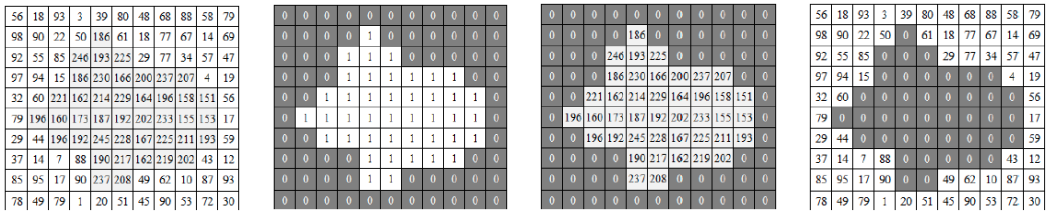
3.2. Extracción de características

Para determinar si una señal detectada en una mamografía corresponde o no una masa, y posteriormente emitir un diagnóstico de si se trata de una lesión benigna o no, se la debe analizar para obtener una serie de propiedades que caracterizan, de forma cuantitiva, cada una de las clases en las que se clasificará posteriormente.

El trabajo de Castro Astudillo [10] presenta un conjunto de características extraídas con el objetivo de representar de forma numérica el contenido visual de las lesiones de masas en mamografías digitales. Un total de 50 características, provenientes del área de computo visual y procesamiento de imágenes, fueron seleccionadas con base en investigaciones previas [6, 12, 28] y revisión de la literatura [48, 50]. Las características contempladas en la presente investigación se dividen, de acuerdo a su naturaleza, en los siguientes grupos:

- **Características de contraste de la señal:** Relacionadas con el nivel de gris de los pixeles que conforman la señal. Se calculan sobre la imagen de la señal segmentada.
- **Características de contraste del fondo:** Relacionadas con el nivel de gris de los pixeles que conforman el fondo de la ventana que contiene la señal. Se calculan sobre la imagen del fondo segmentado.
- **Características de contraste relativo:** Comparan el promedio de gris de la señal con la del fondo.

- **Características de forma:** Describen la forma de la señal. Se calculan sobre la imagen binaria de la señal.
- **Momentos de la secuencia de contorno:** Indican los momentos de forma, promedio y desviación estándar de la distancia al centroide de la señal. También se calculan sobre la imagen binaria de la señal.
- **Momentos geométricos invariantes:** Relacionadas con los momentos invariantes de Hu para reconocimiento de patrones. También se calculan sobre la imagen binaria de la señal.



(a) ROI (b) Imagen binaria (c) Lesión segmentada (d) Fondo segmentado

Figura 3.6: Imágenes sobre las que se realizan los cálculos de características.

3.2.1. Características de contraste de la señal

Son características extraídas de los pixeles que conforman la zona interna a la señal, no considera los pixeles pertenecientes al fondo.

Nivel de gris máximo: Valor máximo de los grises en la señal.

$$máximo = max(imagen) \tag{3.1}$$

Nivel de gris mínimo: Valor mínimo de los grises en la señal.

$$mínimo = min(imagen) \tag{3.2}$$

Mediana del nivel de gris: Mediana de los valores de los grises en la señal. La mediana es el valor ubicado en el centro, luego de ser ordenados.

$$mediana = mediana(imagen) \tag{3.3}$$

Promedio del nivel de gris: Promedio de los valores de los grises de la señal.

$$promedio = \frac{1}{N}\sum_{f=1}^{N_f}\sum_{c=1}^{N_c} imagen(f,c) \tag{3.4}$$

Donde N es el área de la imagen binaria, N_f es el número de filas, N_c es el número de columnas, f es una variable que recorre las filas y c las columnas.

Desviación estándar del nivel de gris: Desviación estándar de los valores de los grises de la señal. Es un buen indicador de nivel de ruido y valores bajos indicarán una supercie suave.

$$\sigma^2 = \frac{1}{N}\sum_{f=1}^{N_f}\sum_{c=1}^{N_c}(imagen(f,c) - \overline{x})^2 \tag{3.5}$$

Donde N es el área de la imagen binaria, N_f es el número de filas, N_c es el número de columnas, f es una variable que recorre las filas y c las columnas y $\overline{x}$ es el promedio de la imagen.

Asimetría del nivel de gris (Skewness): Asimetría de los valores de los grises de la señal. Indica si la cola del histograma se encuentra desviada hacia la derecha (coeciente positivo), en el centro, o hacia la izquierda (coeciente negativo).

$$Sk = \frac{1}{N\sigma^3}\sum_{f=1}^{N_f}\sum_{c=1}^{N_c}(imagen(f,c) - \overline{x})^3 \tag{3.6}$$

Donde N es el área de la imagen binaria, N_f es el número de filas, N_c es el número de columnas, f es una variable que recorre las filas y c las columnas, $\overline{x}$ es el promedio de la imagen y σ es la desviación estándar de la señal.

Kurtosis del nivel de gris: Kurtosis de los valores de los grises de la señal. La kurtosis indica si las colas del histograma tienen una altura superior, igual o inferior a la de una distribución normal. Si el coeciente es negativo se le llamará platicúrtica o platykúrtica y los extremos estarán por debajo de la curva normal. Si el coefciente es igual a cero, se le llamará mesocúrtica o mesokúrtica. Si el coefciente es mayor que cero se le llamará leptocúrtica o leptokúrtica y los extremos estarán por encima la curva normal.

$$K = \frac{1}{N\sigma^4}\sum_{f=1}^{N_f}\sum_{c=1}^{N_c}(imagen(f,c) - \overline{x})^4 - 3 \tag{3.7}$$

Donde N es el área de la imagen binaria, N_f es el número de filas, N_c es el número de columnas, f es una variable que recorre las filas y c las columnas, $\overline{x}$ es el promedio de la imagen y σ es la desviación estándar de la señal.

3.2.2. Características de contraste del fondo

Son características extraídas de los pixeles que conforman el fondo de la señal, no considera los pixeles pertenecientes al interior.

Nivel de gris máximo: Valor máximo de los grises del fondo (ver ecuación 3.1).

Nivel de gris mínimo: Valor mínimo de los grises del fondo (ver ecuación 3.2).

Mediana del nivel de gris: Mediana de los valores de los grises del fondo (ver ecuación 3.3).

Promedio del nivel de gris: Promedio de los valores de los grises del fondo (ver ecuación 3.4).

Desviación estándar del nivel de gris: Desviación estándar de los valores de los grises del fondo (ver ecuación 3.5).

Asimetría del nivel de gris (Skewness): Asimetría de los valores de los grises del fondo (ver ecuación 3.6).

Kurtosis del nivel de gris: Kurtosis de los valores de los grises del fondo (ver ecuación 3.7).

3.2.3. Características de contraste relativo

Son características que intentan encontrar relaciones entre el promedio de gris de la señal con el promedio de gris del fondo, midiendo su contraste.

Contraste absoluto: Promedio de gris en la señal menos el promedio de gris del fondo, dando una medida de contraste.

$$contraste\ absoluto = \overline{x}_{señal} - \overline{x}_{fondo} \tag{3.8}$$

Contraste relativo: Relación entre la diferencia de los promedios de la señal y fondo sobre la suma de los promedios.

$$contraste\ relativo = \frac{\overline{x}_{señal} + \overline{x}_{fondo}}{\overline{x}_{señal} - \overline{x}_{fondo}} \tag{3.9}$$

Contraste proporcional: Relación entre el promedio de la señal sobre el promedio del fondo.

$$contraste\ proporcional = \frac{\overline{x}_{señal}}{\overline{x}_{fondo}} \tag{3.10}$$

3.2.4. Características de forma

Son características geométricas y estructurales que describen la región correspondiente al interior de la señal.

Área: Número de pixeles que ocupa la señal en la región de interés.

$$área = \sum_{f=1}^{N_f} \sum_{c=1}^{N_c} imagen(f, c) \tag{3.11}$$

Donde N es el área de la imagen binaria, N_f es el número de filas, N_c es el número de columnas, f es una variable que recorre las filas y c las columnas y $\overline{x}$ es el promedio de la imagen.

Área convexa: Área que ocupa un polígono convexo que circunscribe a la señal. La figura 3.7 muestra como el polígono de color rojo encierra a la lesión.

Área del fondo: Número de pixeles que ocupa el fondo de la señal.

$$área\ del\ fondo = \sum_{f=1}^{N_f} \sum_{c=1}^{N_c} (1 - imagen(f, c)) \tag{3.12}$$

Donde N es el área de la imagen binaria, N_f es el número de filas, N_c es el número

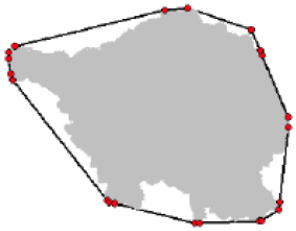
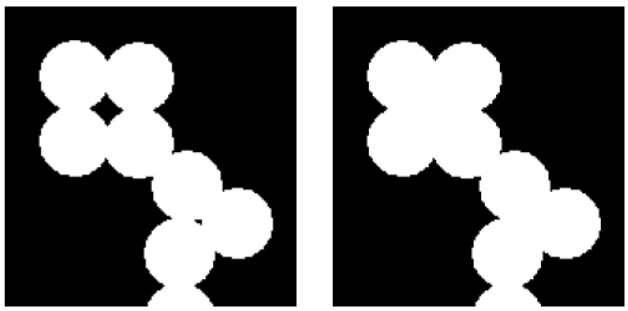


Figura 3.7: Área convexa.

de columnas, f es una variable que recorre las filas y c las columnas y $\overline{x}$ es el promedio de la imagen.

Área rellena: Área que ocupa la señal con todos agujeros rellenos. A partir de la imagen original 3.8(a) se procede a rellenar los agujeros 3.8(b) y luego calcula su área.



(a) Imagen original (b) Imagen rellena

Figura 3.8: Área rellena.

Perímetro: Suma de las distancias entre cada par adyacente de pixeles al rededor del borde de la señal.

Diámetro mayor: Longitud en pixeles del eje mayor de la elipse que tiene los mismos segundos momentos centrales normalizados.

Diámetro menor: Longitud en pixeles del eje menor de la elipse que tiene los mismos segundos momentos centrales normalizados.

Orientación: Ángulo en grados, desde -90 y 90, entre el eje x y y el eje del diámetro mayor.

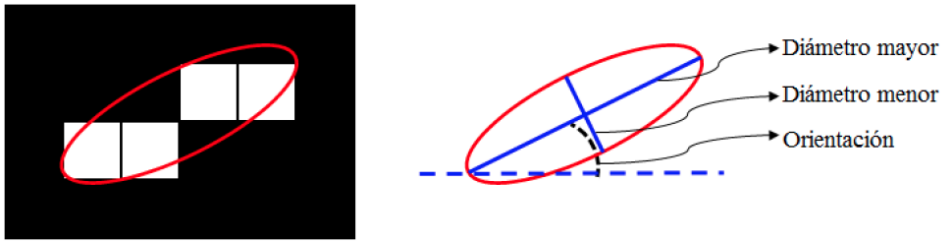


Figura 3.9: Orientación.

Excentricidad: Razón entre la distancia focal, o la distancia entre los focos de la elipse, y su diámetro mayor. Es un indicador de forma, donde valores cercanos a 0 significan que la señal es más redonda.

$$excentricidad = \frac{distancia\ focal}{diámetro\ mayor} \tag{3.13}$$

Número de Euler: Número de objetos en la señal menos el número de agujeros en esos objetos.

$$número\ de\ Euler = número\ de\ objetos - número\ de\ agujeros \tag{3.14}$$

Diámetro circular equivalente: Diámetro de un círculo con área igual a la señal.

$$diámetro\ circular\ equivalente = \sqrt{\frac{4}{\pi} área} \tag{3.15}$$

Solidez: Es la relación entre el área de la señal y el área convexa que lo contiene. Medida de densidad de la señal, donde valores cercanos a 0 significan que el objeto tiene algunos bordes irregulares.

$$solidez = \frac{área}{área\ equivalente} \tag{3.16}$$

(a) 0.90 (b) 0.85 (c) 0.84 (d) 0.75 (e) 1.0

Figura 3.10: Solidez de algunas figuras

Alcance: Relación entre el área de la señal y área de un rectángulo que la abarca.

$$alcance = \frac{área}{área\ rectangulo} \tag{3.17}$$

Factor de forma: Indicador que relaciona el área de la señal con su perímetro y da información sobre su forma. El círculo tiene factor de forma igual a 1, el cuadrado tiene 0,78 y para una bra muy delgada es cercano a 0.

$$factor\ forma = \frac{4\pi * área}{perímetro^2} \tag{3.18}$$

Redondez: Indicador que relaciona el área de la señal con el diámetro máximo y da información sobre cuan circular es la forma. Para el círculo este valor es cercano a 1, en otras formas el valor se incrementa.

$$redondez = \frac{4 * área}{\pi * (diámetro\ máximo)^2} \tag{3.19}$$

Relación de aspecto: Relación entre el diámetro mayor y el diámetro menor de la señal.

$$relación\ de\ aspecto = \frac{diámetro\ mayor}{diámetro\ menor} \tag{3.20}$$

Elongación: Relación entre el diámetro menor y el diámetro mayor de la señal. Si el valor es 1, indica que tiene forma de un cuadrado.

$$enlongación = \frac{diámetro\ menor}{diámetro\ mayor} \tag{3.21}$$

Compacidad: Medida que indica qué tan aglutinada o compacta es la señal. Se puede calcular de las siguientes formas:

$$compacidad_1 = \frac{\sqrt{\frac{4}{\pi} área}}{diámetro\ mayor}$$

$$compacidad_2 = \frac{perímetro^2}{área} \tag{3.22}$$

$$compacidad_3 = \frac{perímetro^2}{4\pi * área}$$

3.2.5. Momentos de Secuencia de Contorno (MSC)

Son características que describen la forma del contorno de la señal. Se los calcula con base en una secuencia de una dimensión que representa la distancia euclidiana entre el centroide y los pixeles del contorno. Dicha secuencia se denomina firma del contorno y se representa de la siguiente forma:

$z(i)$ con $i = 1, 2, 3, \ldots, N$ y $N =$ número de pixeles del contorno.

El n-ésimo descriptor que representa al momento de la secuencia de contorno se calcula de la siguiente forma:

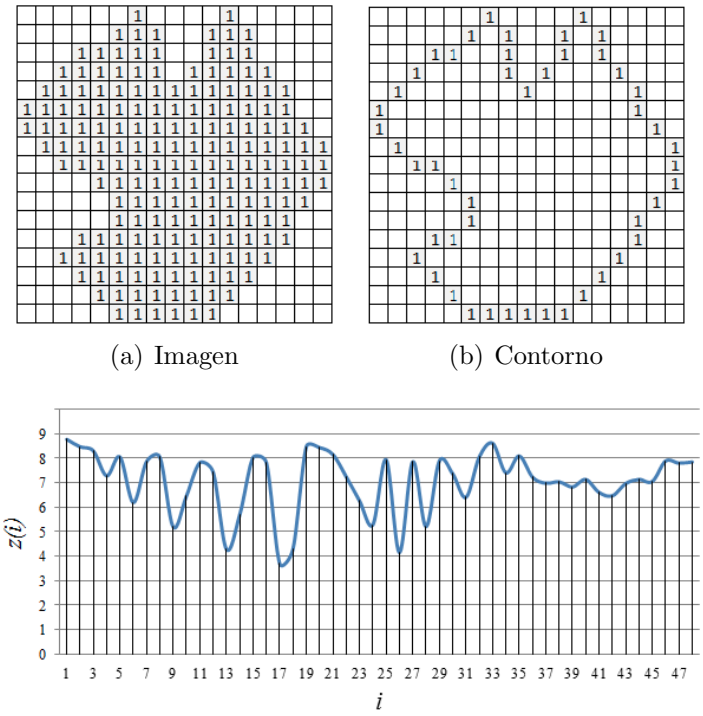
$$F'_n = \frac{\left[\frac{1}{N}\sum_{i=1}^{N}(z(i) - m_1)^n\right]^{\frac{1}{n}}}{m_1}$$

donde:

$$m_1 = \frac{1}{N}\sum_{i=1}^{N} z(i)$$

Shen [51] demuestra estos resultados presentando el cuadro 3.1.

En la presente investigación se utilizan los primeros cuatro descriptores de momentos de la secuencia de contorno, es decir, los F'_n con $n = 1, 2, 3, 4$ ya que son los resultados menos sensibles al ruido. Adicionalmente, se utiliza el promedio y la desviación estándar de las distancias $z(i)$ como características relacionadas a los momentos de la secuencia de contorno.



(a) Imagen (b) Contorno

(c) Firma

Figura 3.11: Firma del contorno de una imagen.

Forma	F_1	F_2	F_3	F'_1	F'_2	F'_3	$F'_3 - F'_1$
Círculo	0.007	0.173	1.929	0.007	0.004	0.008	0.001
Cuadrado	0.108	0.512	2.013	0.108	0.087	0.129	0.021
Rectángulo	0.248	-0.327	1.543	0.248	-0.171	0.277	0.029
Triángulo isósceles	0.305	0.203	2.265	0.305	0.179	0.374	0.069
Triángulo recto	0.371	0.053	1.943	0.371	0.140	0.438	0.067

Cuadro 3.1: Momentos de la secuencia de contorno y descriptores modificados por Shen.

3.2.6. Momentos geométricos invariantes

Son características para reconocimiento de patrones en imágenes en dos dimensiones y su resultado es independiente de la rotación, traslación y cambio de escala.

El momento de orden $(p+q)$ de una imagen $f(x,y)$ se dene como:

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

Para $p, q = 0, 1, 2, \ldots$

El correspondiente momento central se define como:

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

donde

$$\overline{x} = \frac{m_{10}}{m_{00}}$$

$$\overline{y} = \frac{m_{01}}{m_{00}}$$

El momento central normalizado se define como:

$$\eta_{pq} = \frac{\mu_{pq}}{\mu_{00}^{\gamma}}$$

donde

$$\gamma = \frac{p+q}{2} + 1$$

para $p, q = 2, 3, 4, \ldots$

La ecuación 3.23 muestra la forma de calcular los primeros 7 momentos geométricos invariantes, también conocidos como momentos de Hu.

$$
\begin{aligned}
\phi_1 &= \eta_{20} + \eta_{02} \\
\phi_2 &= (\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2 \\
\phi_3 &= (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2 \\
\phi_4 &= (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2 \\
\phi_5 &= (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12}) \left[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2\right] + \\
&\quad (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03}) \left[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2\right] \\
\phi_6 &= (\eta_{20} - \eta_{02}) \left[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2\right] + 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{12} + \eta_{03}) \\
\phi_7 &= (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12}) \left[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2\right] + \\
&\quad (3\eta_{12} - \eta_{30})(\eta_{21} + \eta_{03}) \left[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2\right]
\end{aligned}
\tag{3.23}
$$

3.2.7. Resumen de características

En total se extraen 50 características (ver Cuadro 3.2), de las cuales:

- 7 son de contraste de la señal.
- 7 son de contraste del fondo.
- 3 son de contraste relativo.
- 20 son de forma.
- 6 están relacionadas con momentos de la secuencia de contorno.
- 7 son los primeros momentos invariantes.

	Característica
	Características de contraste de la señal
1	Nivel de gris máximo
2	Nivel de gris mínimo
3	Mediana del nivel de gris
4	Promedio del nivel de gris
5	Desviación estándar del nivel de gris
6	Asimetría del nivel de gris (Skewness)
7	Kurtosis del nivel de gris
	Características de contraste del fondo
8	Nivel de gris máximo
9	Nivel de gris mínimo
10	Mediana del nivel de gris
11	Promedio del nivel de gris
12	Desviación estándar del nivel de gris
13	Asimetría del nivel de gris (Skewness)
14	Kurtosis del nivel de gris
	Características de contraste relativo
15	Contraste absoluto
16	Contraste relativo
17	Contraste proporcional
	Características de forma
18	Área
19	Área convexa
20	Área del fondo
21	Área rellena
22	Perímetro
23	Diámetro mayor
24	Diámetro menor
25	Orientación
26	Excentricidad
27	Número de Euler
28	Diámetro circular equivalente
29	Solidez
30	Alcance
31	Factor de forma
32	Redondez
33	Relación de aspecto
34	Elongación
35	Compacidad 1
36	Compacidad 2
37	Compacidad 3

Cuadro 3.2: Resumen de características

	Característica
	Momentos de secuencia de contorno
38	Momento de secuencia de contorno 1
39	Momento de secuencia de contorno 2
40	Momento de secuencia de contorno 3
41	Momento de secuencia de contorno 4
42	Promedio de radios
43	Desviación estándar de radios
	Momentos invariantes
44	Momento invariante 1
45	Momento invariante 2
46	Momento invariante 3
47	Momento invariante 4
48	Momento invariante 5
49	Momento invariante 6
50	Momento invariante 7

Cuadro 3.2: Resumen de características

Para cada una de las lesiones de cada caso de estudio considerado en la presente investigación, se extrae la totalidad de características listadas en el cuadro 3.2. Estas características representan de forma numérica el contenido de la información visual contenida en la mamografía y será con base en estos datos que se procede a emitir un diagnósticos. En la siguiente sección del documento se aborda la etapa de construcción del modelo mediante la reducción de dimensiones del dominio, descartando aquellas características que en la práctica no aportan poder para discriminar entre las clases en cuestión.

3.3. Selección de características

Una vez determinado el conjunto de características que representan la información contenida en las imágenes y cuyo proceso de extracción se detalla en la sección3.2, se realiza un proceso de selección con el objeto de determinar el subconjunto de características, con el menor tamaño, que logre maximizar el desempeño de un clasificador y así reducir la complejidad computacional asociada a la etapa de clasificación.

Un subconjunto representativo debe ser construido mediante exclusión de entradas que no aportan información útil al modelo al momento de realizar la clasificación de lesiones. En la presente investigación se aplican dos procedimientos, un análisis de la matriz de correlación y un algoritmo de selección de características. El primero intenta buscar pares de características altamente correlacionadas y eliminar uno de cada par. El

segundo esquema considera la totalidad de las características con un bloque completo, realizando una búsqueda en en el espacio de subconjuntos posibles.

Debido a que no se conoce el subconjunto exacto de características que minimizan el error del clasificador, ni la cardinalidad del mismo, es necesario recurrir a un procedimiento de búsqueda que facilite su obtención. Los algoritmos de búsqueda óptima exploran el universo de todas las posibles combinaciones de soluciones, dando como resultado la mejor solución. El total de soluciones a explorar está dado por:

$$\sum_{m=1}^{n} \binom{n}{m} = 2^n \tag{3.24}$$

donde n representa el total de características disponibles y la sumatoria de la parte izquierda de la ecuación 3.24 no es más que el conteo y posterior adición de todos los subconjuntos de tamaño $1, 2, 3, \ldots, n$.

Para las 50 características que se calcularon en la sección anterior, existirían un total de 1,125,899,906,842,623 posibles subconjuntos y, debido a que es un número muy alto, realizar una búsqueda exhaustiva no sería un procedimiento adecuado ya que requeriría gran esfuerzo computacional, tomando en cuenta el número de conjuntos a contemplar y el tiempo de evaluación de cada este proceso conlleva.

Existe otro tipo de algoritmos algoritmos, denominados algoritmos de búsqueda subóptima, que exploran únicamente algunas soluciones, y dan como resultado una solución subóptima que no es necesariamente la mejor, pero reducen el esfuerzo computacional requerido al no explorar todo el conjunto de posibles soluciones.

La figura 3.12 muestra diagrama de flujo del proceso de selección de características:

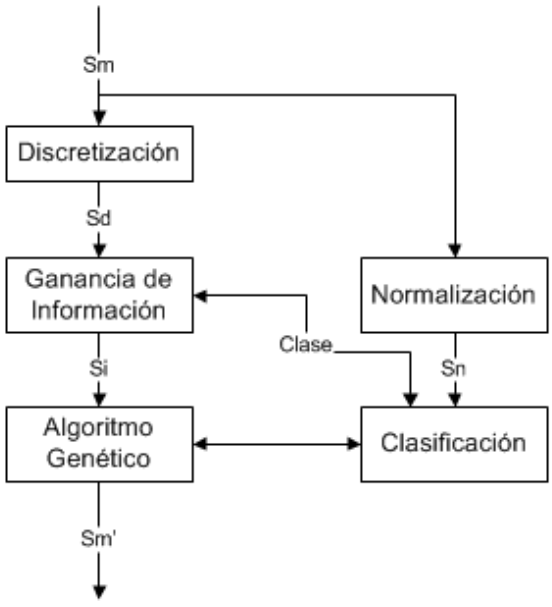


Figura 3.12: Diagrama de flujo del proceso de selección de características

La etapa de selección de características tiene como entrada la lista de características de cada señal (masa) (S_m) y la clase a la que pertenecen (clase). Al final de esta etapa del modelo se obtiene como resultado un subconjunto de características (S'_m), las cuales aportan el mayor poder para discriminar entre masas malignas y benignas. En la etapa de discretización se utiliza un número de intervalos como el utilizado en investigaciones relacionadas [45].

Este proceso es llevado a cabo de forma separada para cada uno de los cuatro algoritmos de clasificación considerados y se consideran diferentes parámetros de operación del algoritmo genético.

3.3.1. Análisis de correlación

El análisis de correlación de un conjunto de variables permite obtener un medida de la relación que existe entra cada una de ellas y se lleva a cabo con el objetivo de identificar aquellas que estén correlacionadas. Una vez determinada la relación entre las variables involucradas en el problema es posible determinar si alguna de ellas puede ser eliminadas, ya que no aporta poder de discriminación con respecto a los valores de clase. En la presente investigación, para eliminar una variable, se consideran los tres criterios que se listan a continuación:

- *Coeficiente de correlación de* 1*:* Se identifica en la matriz de correlación por los valores cercanos a 1. Un valor alto de correlación indica que la forma en que se han calculado las variables guarda algún tipo de relación lineal. Es posible eliminar alguna de las variables.

- *Coeficiente de correlación de* -1*:* Se dice que dos variable tiene un bajo nivel de correlación cuando la matriz de correlación arroja un valor cercano a -1. Al igual que en el caso de un alto valores de correlación, valores cercanos a la unidad (positiva o negativa) indican relación lineal, en este caso con pendientes de signo contrario. Es posible eliminar alguna de las variables.

- *Nivel de correlación cercano a* 1 *o* -1*:* Los valores a considerar en este criterio son aquellos coeficientes de correlación que cumplen con $-1 < correlacion < -0{,}97$ o $0{,}97 < correlacion < 1$. Un nivel de correlación entre estos valores indica correlación estadística entre las variables y se puede proceder a eliminar una de ellas.

Una vez que se ha determinado la correlación entre dos variables se procede a determinar cual de ellas se va a eliminar. Para decidir cual variable conservar dentro del modelo se procede a calcular la *ganancia de información* de dichas variables y se elimina aquella con menor valor.

Para poder obtener la ganancia de información es necesario discretizar las variables. Para ello se utilizó un algoritmo como el descrito en el pseudo-código 2.1 con 10

intervalos para cada variable como se realiza en [45]. Posteriormente se calcula la entropía y finalmente la ganancia de información de cada una de las características con respecto a la clase. Las características son ordenadas de acuerdo a la mayor ganancia de información. De los pares de características correlacionadas detectadas se elimina aquella con la menor ganancia de información.

3.3.2. Algoritmos genéticos

Los algoritmos genéticos se han utilizado para realizar búsquedas subóptimas de subconjuntos de características en problemas diversos [16, 15, 30]. La base de este tipo de algoritmos proviene de una analogía con la evolución de la especies, de tal modo que se afirma que un conjunto de soluciones posibles evoluciona mediante selección natural con el pasar de las generaciones.

La presente investigación utiliza el algoritmo genético (AG) desarrollado por Valenzuela-Rendón [62], denominado algoritmo genético de genes virtuales (vgGA por sus siglas en inglés), con el objetivo de maximizar la sensibilidad y especificidad del clasificador. El vgGA es una generalización de los algoritmos genéticos tradicionales que utilizan cromosomas binarios lineales e implementa las funciones de cruce y mutación como funciones aritméticas sobre números enteros o reales que representa el cromosoma codificado.

Los aspectos más importantes para utilizar un algoritmo genético en un problema de optimización son:

- Representación: La representación es la forma en que se codifican las posibles soluciones en el cromosoma. Como se está utilizando un algoritmo genético para la selección de características, lo que se intenta es incluyendo y excluyendo, de forma guía, características con el objetivo de evaluar su participación en el proceso de clasificación. Bajo esta idea, el cromosoma a utilizar es netamente binario, con alfabeto de 1 y 0, y con una longitud igual 50 características extraídas para cada lesión. Si bien la representación binaria del cromosoma es la más intuitiva, es de notar que al estar utilizando el vgGA, todas las operaciones para evolucionar a los individuos de la población son operaciones aritmeticas en la práctica.

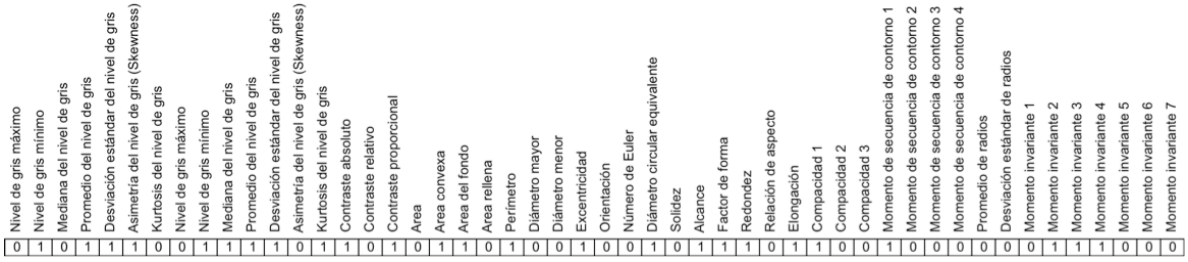
- Función de aptitud: Para evaluar a cada individuo de la población se definió un conjunto de funciones objetivo que involucran la *sensibilidad* y *especificidad* del algoritmo de clasificación que se está utilizando. La justificación proviene del análisis del espacio ROC (véase la sección 2.5.3), donde se pretende maximizar el área bajo la curva. Para cada individuo, es decir para un subconjunto de características, se obtienen las estadísticas de desempeño mediante la validación cruzada de k conjuntos, con una $k = 10$. Los individuos evolucionan hasta alcanzar alcanzar alguno de los criterios de terminacion, determinando un subconjunto de características.

- Método de selección: Como algoritmo de selección se utiliza el de torneo, recomendado en [61, 60]. Se utilizó como alternativa la selección mediante remanente

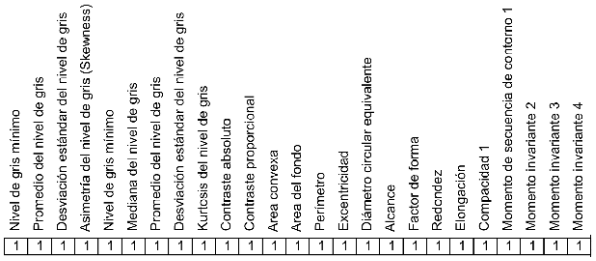
universal estocástico como operador de selección.

- Operador de cruce: como algoritmo de cruce se utiliza el cruce de un punto con parejas aleatorias recomendado en [62, 59].
- Probabilidad de mutación: El valor de la probabilidad de mutación se estableció como se rocomienda en [61], bajo y cercano a 0.1.
- Tamaño de la población: El tamaño de la población varía según el algoritmo de clasificación con el que se trabaja y se adecua según lo expuesto en [30, 61].

De forma general, cuando el cromosoma evoluciona, se analizan sus valores para determinar el subconjunto de características que se seleccionan y que sirven como entrada a los algoritmos de clasificación. Si el valor en la posición i del cromosoma es 1, la característica que se ubica en la posición i es seleccionada, como se observa en la figura 3.13.



(a) Cromosoma de máxima longitud



(b) Características seleccionadas

Figura 3.13: Cromosoma del algoritmo genético

La figura 3.13(b) muestra las características consideradas en el cromosoma de la figura 3.13(a).

Se realiza una segunda búsqueda con algoritmo genético utilizando los mismo parámetros pero con un cromosoma de longitud igual a 50 menos el número de características excluídas por el análisis de correlación.

3.4. Clasificación

La última etapa de modelo de solución propuesta en esta investigación es la clasificación automática de masas tumorales. Previo a esta última etapa la información visual relevante contenida en las mamografías seleccionadas de la base de datos DDSM ha sido convertida a ejemplos numéricos que pueden ser tratados y manipulados por algoritmos computacionales así también se ha determinado el conjunto de entradas que maximiza el rendimiento de cada uno de los algoritmos de clasificación seleccionados: Redes Neuronales (NN), Máquinas de Vector de Soporte (SVM), Análisis de Discriminante Lineal (LDA) y Regresión Logística (RL).

En la etapa de clasificación se determina si una masa es maligna o beninga, de acuerdo al modelo interno definido para los algoritmos antes listados. Antes de realizar la clasificación de los ejemplos es recomendable llevar a cabo un proceso de normalización de datos [10, 45].

3.4.1. Normalización de valores

Debido a que los valores de las características pueden ser cualquier número positivo o negativo, es necesario realizar una normalización de los datos de entrada y de salida en un rango entre $[N_{min}, N_{max}]$ para que no afecte el comportamiento del clasificador.

El método más comúnmente utilizado es el *escalamiento lineal simple* [45], en el que se calculan los datos normalizados en función del valor máximo y el valor mínimo de cada característica. Su fórmula es:

$$N = N_{max} - (N_{max} - N_{min}) * \frac{Max - valor}{Max - Min} \tag{3.25}$$

En la presente investigación se normalizó los valores de los ejemplos a un rango que proviene de investigaciones previas realizadas con un clasificador neuronal [28] donde la normalización de valores se lleva al espacio $[-1, 1]$.

3.4.2. Red neuronal (NN)

Las redes neuronales (*neural network*) son un algoritmo de aprendizaje supervisado que se usa para clasificación. En el presente trabajo se utiliza una re neuronal para decidir si una masa contenida en una mamografía es maligna o beninga a partir de las características extraidas de la lesión. Se utiliza una red neuronal 3 capas que aplica como algoritmo de aprendizaje la retropropagación y cuya arquitectura ha sido también utilizada en trabajos similares [10, 45] y que se apega a lo establecido en el teorema de Kolgomorov (ver sección 2.4.1).

Como entrada de la red se define un número de neuronas n igual al número de características seleccionadas en la etapa anterior, la de selección de caracterísitcas. Con base en el teorema de Kolgomorov, el número de neuronas de la capa intermedia es

$2n + 1$ [36]. La capa de salida esta formada por una única unidad, debido a que es suficiente para el proceso de clasificación binaria entre masas malignas y benignas. La función de transferencia entre los nodos es la tangente hiperbólica sigmoidal y la función de medida del error es el error cuadrático medio, como lo utilizado en [10].

La red de retropropagación se entrenó durante 500 épocas.

3.4.3. Máquina de vector de soporte (SVM)

Las *máquinas de vector de soporte* son un modelo de aprendizaje supervisado y se han convertido en un algoritmo muy utilizado para el reconocimiento de patrones y clasificación binaria. La idea central consiste en determinar un hiperplano que maximiza la distancia a los ejemplos más cercanos, o *vectores de soporte*. El análisis de donde deriva el modelo de vectores de soporte se detalla en la sección 2.4.2.

En la presente investigación se implementa un tipo de máquina de vectores de soporte con escalamiento a modo de alcanzar desviación de los datos igual a la unidad. Como función de kernel se utiliza *Gaussian Radial Basis Function* con $\sigma = 2$ como se ha previamente utilizado en [20]. Como método para determinar el hiperplano se utiliza la programación cuadrática *Quadratic Programming* para optimizar la separación entre el hiperplano y los vectores de soporte.

3.4.4. Análisis de discriminante (LDA)

El análisis de discriminante es un método utilizado en estadística, reconocimiento de patrones, aprendizaje automático para encontrar una combinación lineal de características que permita separar dos o más clases de objetos. En este tipo de análisis se cuenta con un conjunto de variables continuas independientes y una variable nominal para la clase. El análisis tiene como base la suposición que las distribuciones condicionales de probabilidad de las clases son distribuciones normales 2.4.3.

En el presente trabajo de investigación se utiliza el análisis discriminante para predecir la clase de los ejemplos extraídos de la base DDSM. Se construye un modelo LDA con función discriminante *diaglinear*, la cual es una función lineal estima una distribución normal de probabilidad para las cada una de las clases, positiva (masa maligan) y negativa (masa benigna), mediante la estimación de la varianza (*diagonal covariance matrix estimate*). En el modelo no se consideran las probabilidades *a priri* para clase.

3.4.5. Regresión Logística (RL)

La regresión logística es uno de los modelos lineales generalizados que se usa para predecir el resultado de una variable nominal, como la el diagnóstico de masas en mamografías, en función de variables continuas independientes, mediante el uso de una función logística. Es muy utilizado en la predicción médica de diferentes enfermedades [29].

La presente investigación implementa un modelo de regresión logística para la clasificación binaria de lesiones de masas tumorales en mamografías. Los coeficientes de la regresión se estimas mediante un algoritmo de regresión logística multinomial y luego son utilizados para predecir la probabilidad de pertenecer a una u otra clase [5, 14, 65]. En el modelo se considera la clase *nominal* y como función de enlace se considera la tradicional *logit* (inverso de la función sigmoide) para estimar los predictores del modelo y posteriormente maximizar la *verosimilitud* de los mismos.

3.4.6. Validación

Para obtener las métricas de desempeño de los algoritmos de clasificación red neuronal NN, máquina de vector de soporte SVM, análisis discriminante LDA y regresión logística LR, se utiliza nuevamente la validación cruzada de 10 conjuntos. Se construye cada clasificador con el modelo determinado en la selección de características y posteriormente se reportan los resultados obtenidos.

3.5. Resumen

El modelo de solución presentado en este capítulo describe el proceso para selección de características y evaluación del desempeño de algoritmos de clasificación en masas detectadas en mamografías digitales. El modelo se divide en las siguiente etapas: *pre-procesamiento*, *segmentación* y *extracción de características*, así como la *reducción de dimensiones* y *clasificación*.

La primera parte del modelo se enfoca en la contrucción de un conjunto de ejemplos sobre masas detectadas en mamografías. Para ellos se utiliza la base de datos *Digital Database for Screening Mamography DDSM*. Para los casos de masas seleccionados, se toma la mamografía, se pre-procesa mediante un filtro de mediana, con el objeto de eliminar el ruido contenido en ella. Posteriormente se hace uso de la información proporcionada por la base DDSM para delimitar la lesión en cuestión y posteriormente segmentar la región de interés que contiene a la masa. Una vez obtenida esta región de interés se extrae un conjunto de características que representan la información visual contenida en la mamografía y que representa a la masa de forma cuantitativa. Cada masa se encuentra ahora representada por un total de 50 características numéricas.

Ya que se cuenta con un conjunto de ejemplos que se puede ser manipulado por algoritmos computacionales se procede a reducir la cantidad de las entradas de los algoritmos de clasificación. El objetivo es descartar del modelo aquellas características que no aporten poder para discriminar entre masas malignas y benignas. Este proceso se lleva a cabo mediante un enfoque híbrido que utiliza el análisis discriminante para eliminar características correlacionadas y posteriormente una búsqueda con un algoritmo genético para determinar el conjunto de menor tamaño con el mejor rendimiento para construir un clasificador. Se utilizan cuatro algoritmos de clasificación y se distingue

entre subconjuntos con base en el rendimiento de cada clasificador construido y su validación cruzada con 10 conjuntos.

La última etapa del modelo evalúa los resultados obtenidos y entrega medidas de desempeño de los algoritmos de clasificación. Las métricas consideradas en esta investigación son el éxito, sensibilidad, especificidad y error.

Capítulo 4

Experimentos y Resultados

El presente capítulo detalla los experimentos realizados para sustentar la investigación así como los resultados obtenidos en cada uno de ellos. En la primera parte del capítulo se abordan los aspectos relacionados con la construcción de un conjunto de ejemplos a partir de la base de datos DDSM y que fueron utilizados para la construcción y validación del modelo propuesto. Posteriormente se describen los experimentos realizados y se muestran los resultados obtenidos para cada una de las fases del modelo de solución. Se incluye un análisis de

4.1. Plataforma experimental

El software utilizado para implementar el procedimiento planteado en los capítulos anteriores fue desarrollado en MATLAB Release 12a 64 bits. Se hizo uso intensivo de las herramientas disponibles en el software (Tools Boxs), tales como el Image Processing Toolbox, Neural Network Toolbox y el Bioinformatics Toolbox; además se desarrollaron un conjunto de programas para implementar los requerimientos del sistema. Las corridas se hicieron en una computadora con un procesador Intel Xeon CPU de 4 GB RAM y 1 TB de disco duro. El sistema operativo utilizado fue el Microsoft Windows 7 64 bits Enterprise, Service Pack 1 y no fue necesario el uso de procesamiento en paralelo.

4.2. DDSM

La DDSM [26, 27] está organizada en casos y volúmenes. Un volumen es una colección de casos, los cuales han sido agrupados para facilitar la distribución de los mismos. La base de datos está constituida por 2632 casos de estudio, cada uno de los cuales contiene dos imágenes de cada mama, que corresponden a las vistas cráneo-caudal (CC) y medio-lateral oblicua (MLO) de cada paciente. La base de datos agrupa un total de 10528 mamografías, que fueron digitalizadas a partir de las placas radiográficas originales mediante el uso de cuatro escáneres diferentes que se describen en la tabla 2.8.

La base de datos DDSM ha sido utilizada en diversas investigaciones que abordan la temática de apoyo al diagnóstico de cáncer de mama. Concretamente en cuanto lesiones tumorales se resaltan [11, 21, 64]

4.2.1. Resumen de la base de datos

La base de datos proporciona, para cada mamografía donde se haya detectado una lesion de masa, un archivo anexo donde se detalla la información visual de la lesión contenida en la imagen, especificando los siguientes campos:

- Total de anormalidad (Total abnormality): indica el número de lesiones detectadas en dicha imagen.I
- Anormalidad (Abnormality): En caso que se hayan encontrado más de un hallazgo en la imagen, indica el ordinal de la lesión que se está describiendo.
- Tipo de lesión (Lesion type): Se especifica si la lesión es una masa.

  Forma de la masa (Shape): Esta puede tomar alguno de los valores: redonda (round), ovalada (oval), lobulada (lobulated) e irregular (irregular) o distorsión de arquitectura (architectural distorsion).

  Margen (Margins): Puede tomar cualquier de los siguientes valores: circunscrita (circunscribed), microlobulado (microlobulated), ensombrecido (obscured), mal definido (ill-defined) y espiculado (spiculated).
- Nivel de apreciación (Assessment): Puede tomar un valor de 1 a 5 y corresponde con la especificación del BI-RADS.
- Sutileza (Sublety): Puede tomar cualquier valor de 1 a 5 donde 1 es valor más débil y 5 indica que la lesión es obviamente apreciable,
- Patología (Patology): Se indica si la lesión es maligna (M) o benigna (B).
- Total de contornos (Total outlines): Toma un valor numérico y representa el número de contornos que encierran a las masas encontradas en la imagen.
- Contorno (Outline): Cadena de caracteres que especifica la secuencia de puntos que se debe seguir para determinar el contorno que encierra a la lesión

Esta información se encuentra contenida en el archivo que se especifica con detalle en la sección 2.6.2.

4.2.2. Análisis de la base de datos

Se efectuó un análisis de los casos contenidos en la base de datos DDSM con el objetivo de identificar el conjunto de datos que represente a la lesión en estudio. Las consideraciones realizadas se listan a continuación:

- Se agregó la columna (Mama) que indica si la imagen corresponde a la mama izquierda (L) o mama derecha (R), tomando en cuenta lo que indica la descripción de la base de datos.
- Se aumentó la columna (Vista) que indica si la mamografía corresponde a la proyección Cráneo-Caudal (CC) o medio lateral oblicua (MLO), tomando en cuenta lo que indica la descripción de la base de datos.
- Se seleccionaron los casos con lesiones contenidas en un único contorno.
- Los casos B_3020 y B_3411 que son descritos en la base de datos con un borde CIRC-ILLD y ILLD-SPIC, respectivamente, se presentan como ILLD, por ser la forma más general entre ambas anotaciones.
- Los casos B_3001, B_3022, B_3109, B_3134, B_3373, B_3377 y B_3508 que se describen como IRRE-ARCH, se presentan como ARCH, por ser la forma más general entre ambas anotaciones.

Se considera un total de 180 mamografías provenientes de 119 casos donde se detectó una masa delimitada por único borde, de las cuales 95 tienen un diagnóstico de benigno y 85 maligno. 99 mamografías corresponden a la vista cráneo-caudal CC y de esas 51 tienen un diagnostico maligno. Las 81 mamografías restantes corresponde a la vista medio-lateral oblicua MLO, 34 contienen una masa con diagnóstico de cáncer.

De las 85 masas diagnosticadas como malignas, 72 corresponden a masas con margen espiculado. En este caso se consideraron únicamente casos donde las masas detectadas tenías un nivel de apreciación superior o igual a 4.

La tabla 4.1 muestra la lista de casos considerados en este estudio, así como la información extraída de la base de datos DDSM.

Caso	Forma	Margen	Apreciación	Dx	Mama	Vista
A_1093	IRRE	SPIC	5	M	R	CC
A_1114	OVAL	SPIC	4	M	L	CC
A_1114	OVAL	SPIC	4	M	L	MLO
A_1118	IRRE	SPIC	4	M	R	MLO
A_1122	IRRE	SPIC	5	M	L	MLO
A_1134	IRRE	SPIC	4	M	R	MLO
A_1147	IRRE	SPIC	5	M	L	CC

Cuadro 4.1: Resumen de la base de datos DDSM

Caso	Forma	Margen	Apreciación	Dx	Mama	Vista
A_1149	IRRE	SPIC	4	M	L	CC
A_1160	IRRE	SPIC	4	M	R	MLO
A_1163	IRRE	SPIC	5	M	R	MLO
A_1168	IRRE	SPIC	4	M	L	CC
A_1168	IRRE	SPIC	4	M	L	MLO
A_1170	LOBU	CIRC	4	B	R	CC
A_1170	LOBU	CIRC	4	B	R	MLO
A_1177	LOBU	OBSC	3	B	R	CC
A_1177	LOBU	OBSC	3	B	R	MLO
A_1224	IRRE	SPIC	4	M	L	CC
A_1224	IRRE	SPIC	4	M	L	MLO
A_1234	IRRE	SPIC	5	M	R	CC
A_1234	IRRE	SPIC	5	M	R	MLO
A_1236	IRRE	SPIC	5	M	R	CC
A_1236	IRRE	SPIC	5	M	R	MLO
A_1237	IRRE	SPIC	5	M	L	CC
A_1237	IRRE	SPIC	5	M	L	MLO
A_1264	OVAL	CIRC	3	B	R	CC
A_1264	OVAL	CIRC	3	B	R	MLO
A_1266	LOBU	ILLD	3	B	L	CC
A_1266	LOBU	ILLD	3	B	L	MLO
A_1267	OVAL	CIRC	4	B	R	CC
A_1323	OVAL	OBSC-ILLD	4	B	L	CC
A_1323	OVAL	OBSC-ILLD	4	B	L	MLO
A_1346	OVAL	ILLD	4	B	L	CC
A_1346	OVAL	ILLD	4	B	L	MLO
A_1347	IRRE	ILLD	4	B	L	CC
A_1347	IRRE	ILLD	4	B	L	MLO
A_1348	ROUND	CIRC	4	B	L	CC
A_1348	ROUND	CIRC	4	B	L	MLO
A_1354	OVAL	CIRC	4	B	R	CC
A_1354	OVAL	CIRC	4	B	R	MLO
A_1403	IRRE	SPIC	5	M	R	CC
A_1403	IRRE	SPIC	5	M	R	MLO
A_1416	IRRE	SPIC	5	M	R	CC
A_1416	IRRE	SPIC	5	M	R	MLO
A_1417	IRRE	SPIC	5	M	R	MLO
A_1510	IRRE	SPIC	5	M	L	CC
A_1510	IRRE	SPIC	5	M	L	MLO
A_1520	IRRE	SPIC	4	M	R	CC
A_1520	IRRE	SPIC	4	M	R	MLO
A_1622	IRRE	ILLD	4	M	R	CC

Cuadro 4.1: Resumen de la base de datos DDSM

Caso	Forma	Margen	Apreciación	Dx	Mama	Vista
A_1622	IRRE	ILLD	4	M	R	MLO
A_1642	IRRE	SPIC	5	M	R	MLO
A_1700	IRRE	SPIC	5	M	R	MLO
A_1701	IRRE	SPIC	5	M	L	MLO
A_1821	IRRE	SPIC	5	M	L	MLO
A_1827	ROUND	ILLD	4	M	R	CC
A_1827	ROUND	ILLD	4	M	R	MLO
B_3001	IRRE-ARCH	SPIC	5	M	L	MLO
B_3012	IRRE	SPIC	4	M	R	MLO
B_3016	IRRE	SPIC	4	M	R	MLO
B_3017	LOBU	ILLD	4	M	L	MLO
B_3018	LOBU	ILLD	4	M	L	CC
B_3018	LOBU	ILLD	4	M	L	MLO
B_3020	LOBU	CIRC-ILLD	4	M	R	MLO
B_3022	IRRE-ARCH	SPIC	5	M	R	CC
B_3032	ROUND	ILLD	4	M	L	MLO
B_3074	IRRE	SPIC	4	M	R	CC
B_3084	LOBU-IRRE	SPIC	5	M	R	CC
B_3091	IRRE	ILLD	4	B	L	CC
B_3091	OVAL	ILLD	4	B	L	MLO
B_3093	ARCH	ILLD-SPIC	4	B	L	MLO
B_3096	IRRE	OBSC	3	B	L	CC
B_3096	IRRE	ILLD	3	B	L	MLO
B_3096	IRRE	SPIC	4	B	R	MLO
B_3097	ROUND	CIRC-OBSC	3	B	R	CC
B_3097	OVAL	CIRC-OBSC	3	B	R	MLO
B_3099	IRRE	ILLD	3	B	L	CC
B_3099	IRRE	ILLD	4	B	L	MLO
B_3100	ROUND	MICR	4	B	L	CC
B_3100	IRRE	OBSC-ILLD	3	B	L	MLO
B_3102	OVAL	CIRC	4	B	L	CC
B_3102	OVAL	CIRC	4	B	L	MLO
B_3102	LOBU	CIRC-ILLD	4	B	R	CC
B_3102	LOBU	CIRC-ILLD	4	B	R	MLO
B_3109	IRRE-ARCH	SPIC	4	M	R	CC
B_3113	ROUND	CIRC	3	B	R	CC
B_3113	ROUND	CIRC-ILLD	4	B	R	MLO
B_3114	LOBU	CIRC	4	B	L	CC
B_3118	OVAL	CIRC	3	B	L	CC
B_3122	ROUND	MICR	4	B	L	CC
B_3122	OVAL	ILLD	3	B	L	MLO
B_3126	ROUND	CIRC	3	B	R	MLO

Cuadro 4.1: Resumen de la base de datos DDSM

Caso	Forma	Margen	Apreciación	Dx	Mama	Vista
B_3128	IRRE	ILLD-SPIC	4	B	R	CC
B_3128	IRRE	ILLD-SPIC	4	B	R	MLO
B_3129	IRRE	OBSC-ILLD	4	B	R	CC
B_3129	IRRE	ILLD-SPIC	4	B	R	MLO
B_3132	LOBU	CIRC-ILLD	4	B	R	CC
B_3132	OVAL	CIRC	3	B	R	MLO
B_3134	IRRE-ARCH	SPIC	5	M	R	CC
B_3140	OVAL	CIRC	4	B	R	CC
B_3142	OVAL	CIRC-OBSC-ILLD	4	B	R	CC
B_3143	ROUND	CIRC	3	B	L	CC
B_3143	ROUND	CIRC	3	B	L	MLO
B_3144	OVAL	CIRC-ILLD	4	B	R	CC
B_3144	IRRE	ILLD	4	B	R	MLO
B_3146	LOBU	CIRC	4	B	R	CC
B_3146	IRRE	MICR-ILLD	4	B	R	MLO
B_3151	LOBU	CIRC-OBSC	4	B	L	CC
B_3151	LOBU	CIRC	4	B	L	MLO
B_3153	IRRE	ILLD	4	B	L	CC
B_3153	IRRE	ILLD-SPIC	4	B	L	MLO
B_3154	OVAL	CIRC	3	B	L	CC
B_3154	OVAL	CIRC	3	B	L	MLO
B_3156	OVAL	CIRC-OBSC-ILLD	4	B	R	CC
B_3156	OVAL	CIRC-OBSC-ILLD	4	B	R	MLO
B_3158	LOBU	CIRC	3	B	L	CC
B_3158	LOBU	CIRC-OBSC	3	B	L	MLO
B_3186	OVAL	CIRC	2	B	L	MLO
B_3373	IRRE-ARCH	SPIC	5	M	R	CC
B_3377	IRRE-ARCH	SPIC	5	M	R	CC
B_3411	IRRE	ILLD-SPIC	5	M	R	CC
B_3499	IRRE	SPIC	5	M	R	CC
B_3508	IRRE-ARCH	SPIC	5	M	R	CC
C_0001	IRRE	SPIC	5	M	R	CC
C_0003	IRRE	SPIC	5	M	R	CC
C_0004	IRRE	SPIC	5	M	R	CC
C_0006	IRRE	SPIC	5	M	R	CC
C_0009	IRRE	SPIC	5	M	R	CC
C_0014	IRRE	MICR	5	M	R	MLO
C_0015	IRRE	SPIC	5	M	R	CC
C_0015	IRRE	SPIC	5	M	R	MLO
C_0016	IRRE	SPIC	5	M	R	CC
C_0016	IRRE	SPIC	5	M	R	MLO
C_0017	LOBU	ILLD	4	M	L	MLO

Cuadro 4.1: Resumen de la base de datos DDSM

Caso	Forma	Margen	Apreciación	Dx	Mama	Vista
C_0019	OVAL	CIRC	2	B	L	MLO
C_0019	LOBU	CIRC	5	M	R	MLO
C_0029	OVAL	ILLD	3	B	L	CC
C_0029	OVAL	ILLD	3	B	L	MLO
C_0031	IRRE	SPIC	5	M	R	CC
C_0033	OVAL	MICR	4	B	R	CC
C_0033	LOBU	MICR	4	B	R	MLO
C_0034	IRRE	SPIC	5	M	R	CC
C_0061	IRRE	SPIC	5	M	R	CC
C_0064	IRRE	SPIC	5	M	R	CC
C_0069	IRRE	SPIC	5	M	R	CC
C_0085	IRRE	SPIC	5	M	R	CC
C_0102	IRRE	SPIC	5	M	R	CC
C_0121	IRRE	SPIC	5	M	R	CC
C_0156	IRRE	SPIC	5	M	R	CC
C_0181	IRRE	SPIC	5	M	R	CC
C_0186	IRRE	SPIC	5	M	R	CC
C_0194	IRRE	SPIC	5	M	R	CC
C_0201	IRRE	SPIC	5	M	R	CC
C_0217	ROUND	CIRC	3	B	R	CC
C_0230	IRRE	SPIC	5	M	R	CC
C_0235	OVAL	MICR	4	B	R	CC
C_0235	OVAL	MICR	4	B	R	MLO
C_0237	ROUND	CIRC	4	B	L	CC
C_0237	OVAL	CIRC	4	B	L	MLO
C_0240	OVAL	ILLD	4	B	R	CC
C_0240	OVAL	CIRC	4	B	R	MLO
C_0241	IRRE	ILLD	4	B	L	CC
C_0241	IRRE	ILLD	4	B	L	MLO
C_0243	OVAL	CIRC	4	B	R	CC
C_0243	OVAL	MICR	4	B	R	MLO
C_0245	OVAL	CIRC	4	B	R	CC
C_0245	OVAL	CIRC	4	B	R	MLO
C_0247	LOBU	ILLD	3	B	R	CC
C_0247	LOBU	CIRC	3	B	R	MLO
C_0248	OVAL	ILLD	4	B	R	CC
C_0248	LOBU	MICR	4	B	R	MLO
C_0249	LOBU	CIRC	4	B	L	CC
C_0249	LOBU	CIRC	4	B	L	MLO
C_0249	LOBU	CIRC	4	B	R	CC
C_0249	LOBU	CIRC	4	B	R	MLO
C_0250	ROUND	CIRC	4	B	R	CC

Cuadro 4.1: Resumen de la base de datos DDSM

Caso	Forma	Margen	Apreciación	Dx	Mama	Vista
C_0250	OVAL	CIRC	4	B	R	MLO
C_0339	ROUND	SPIC	5	M	R	CC
C_0340	IRRE	SPIC	5	M	R	CC
D_4174	IRRE	SPIC	5	M	R	CC
D_4185	IRRE	SPIC	5	M	R	CC

Cuadro 4.1: Resumen de la base de datos DDSM

4.3. Extracción de características

Las 50 características descritas en la 3.2 se extraen a la imágenes resultantes después del proceso de segmentación3.1. Un resumen de los valores calculados, indicando los estadísticos de mínimo valor (Min.), máximo valor (Max.), media (Med.) y desviación estándar(Des.), se presenta en el cuadro 4.2 para el conjunto de ejemplos seleccionados de la base DDSM 4.1.

	Característica	Min	Max	Prom	Des
	Características de contraste de la señal				
1	Nivel de gris máximo	2542	61888	50231.5	11956.35
2	Nivel de gris mínimo	96	46176	31107.31	10721.04
3	Mediana del nivel de gris	1662	56480	41444.02	10516.90
4	Promedio del nivel de gris	1897.9	53766	41130.13	10408.36
5	Desviación estándar del nivel de gris	112.56	11690	3485.873	2170.455
6	Asimetría del nivel de gris (Skewness)	-2.5318	1.501	-0.19556	0.6332
7	Kurtosis del nivel de gris	1.4903	13.533	2.7893	1.2145
	Características de contraste del fondo				
8	Nivel de gris máximo	54592	65520	64806.81	2382.388
9	Nivel de gris mínimo	1	12208	6593.916	4772.524
10	Mediana del nivel de gris	88	41088	19747.91	8874.467
11	Promedio del nivel de gris	805.7	34881	23226.32	6112.806
12	Desviación estándar del nivel de gris	904.19	20738	13628.14	3765.363
13	Asimetría del nivel de gris (Skewness)	-0.62703	6.9361	0.64522	0.8015
14	Kurtosis del nivel de gris	1.1863	89.254	2.9498	6.7139
	Características de contraste relativo				
15	Contraste absoluto	-2269.2	33855	17903.80	7266.803
16	Contraste relativo	-0.03717	0.48688	0.27504	0.088421
17	Contraste proporcional	0.92832	2.8977	1.8	0.34533

Cuadro 4.2: Resumen de características

	Característica	Min	Max	Prom	Des
	Características de forma				
18	Área	10968	3180000	223819.7	312653.9
19	Área convexa	12592	3450000	233350.8	329402.4
20	Área del fondo	5464	2550000	88382.19	203147.8
21	Área rellena	10968	3180000	223819.9	312653.8
22	Perímetro	494.86	8976.4	1748.891	1044.520
23	Diámetro mayor	152.13	3139.6	548.8344	349.6359
24	Diámetro menor	90.098	1386.6	407.1553	223.6844
25	Orientación	0.19369	0.91617	0.61634	0.15778
26	Excentricidad	-89.364	89.91	0.12187	57.7179
27	Número de Euler	-2	1	0.95556	0.29563
28	Diámetro circular equivalente	118.17	2011.6	463.6301	265.3599
29	Solidez	0.60819	0.99522	0.94702	0.068177
30	Alcance	0.39684	0.85546	0.72081	0.093498
31	Factor de forma	0.14255	0.88336	0.71722	0.14447
32	Redondez	0.34057	0.97289	0.7402	0.13898
33	Relación de aspecto	1.0193	2.4951	1.3578	0.27069
34	Elongación	0.40078	0.98106	0.76104	0.12758
35	Compacidad 1	0.58359	0.98635	0.85623	0.084317
36	Compacidad 2	14.226	88.156	18.9278	8.0805
37	Compacidad 3	1.132	7.0153	1.5062	0.64303
	Momentos de secuencia de contorno				
38	Momento de secuencia de contorno 1	0.017128	0.37287	0.12962	0.073883
39	Momento de secuencia de contorno 2	0.006828	0.27061	0.074924	0.053659
40	Momento de secuencia de contorno 3	0.021718	0.45319	0.15696	0.090657
41	Momento de secuencia de contorno 4	0.016386	0.42527	0.12748	0.083716
42	Promedio de radios	65.073	981.25	234.3003	133.4589
43	Desviación estándar de radios	4.1972	365.91	30.5625	35.6196
	Momentos invariantes				
44	Momento invariante 1	0.15935	0.2913	0.1747	0.019914
45	Momento invariante 2	9.42e-6	0.03537	0.003689	0.005490
46	Momento invariante 3	2.01e-7	0.012874	0.000437	0.001344
47	Momento invariante 4	1.17e-10	0.003398	6.3518e-5	0.000312
48	Momento invariante 5	-2.88e-8	4.25e-5	2.9582e-7	3.1852e-6
49	Momento invariante 6	-9.91e-6	0.000588	7.1464e-6	4.948e-5
50	Momento invariante 7	-4.41e-7	7.53e-6	6.0555e-8	6.0274e-7

Cuadro 4.2: Resumen de valores de características extraídas

La siguiente etapa del modelo de solución se encarga de reducir el conjunto de características que se suministran a los algoritmos de clasificación para que realicen su

labor.

4.4. Selección de características

En esta sección de abordan dos enfoques diferentes de reducción de dimensiones: el *análisis de correlación*, cuyo origen es la estadística; y el *modelo de envoltura* mediante algoritmos genéticos para la exploración subóptima del espacio de búsqueda. Para ambos enfoques se detallan los aspectos relacionados y se presentan los resultados obtenidos en la experimentación.

4.4.1. Análisis de correlación

Se realizó el análisis de correlación de la totalidad de características extraídas. La información se presenta dividida de acuerdo a la naturaleza de los grupos de características, una tabla por grupo. La información del grupo de características de forma de la masa se presenta en 3 tablas para una mejor compresión de los resultados, aunque el análisis se realizó sobre la matriz completa.

Características contraste de la señal:
El cuadro 4.3 muestra la matriz de correlación para las características contraste de la señal. Los valores de correlación en el nivel máximo de gris, nivel mínimo de gris, promedio del nivel de gris y mediana del nivel de gris son altos, obteniendo un valor mínimo de 0.889 entre cualquier par de estas características. Con base en los criterios especificados en la sección 3.3.1 se determino que existe correlación entre promedio del nivel de gris y mediana del nivel de gris al obtener un nivel de correlación de 0.996, muy cercano a la unidad. Una de estas características puede ser eliminada.

	Max	Min	Mediana	Prom	Des	Skewness	Kurtosis
Max		0.655	0.918	0.919	0.414	-0.14	0.101
Min	0.655		0.807	0.837	-0.33	0.045	0.026
Mediana	0.918	0.807		0.996	0.214	-0.28	0.084
Prom	0.919	0.837	0.996		0.173	-0.23	0.084
Des	0.414	-0.33	0.214	0.173		-0.24	-0.09
Skewness	-0.14	0.045	-0.28	-0.23	-0.24		-0.19
Kurtosis	0.101	0.026	0.084	0.084	-0.09	-0.19	

Cuadro 4.3: Matriz de correlación para características contraste de la señal

Del análisis de ganancia de información se tiene que la mediana del nivel de gris tiene una ganancia de información al **promedio del nivel de gris**, por lo que este último es eliminado del conjuntos features.

Características contraste del fondo:
El cuadro 4.4 presenta los niveles de correlación para las características contraste del fondo. Las características con mayor nivel de correlación son el promedio del nivel de

gris y la mediana del nivel de gris con un nivel de 0.802. No se detectó correlación entre variables de este grupo de características. No se eliminó ninguna característica de este grupo.

	Max	Min	Mediana	Prom	Des	Skewness	Kurtosis
Max		0.341	0.439	0.759	0.675	0.052	-0.05
Min	0.341		-0.06	0.257	-0.18	0.236	-0.06
Mediana	0.439	-0.06		0.802	0.569	-0.54	-0.23
Prom	0.759	0.257	0.802		0.758	-0.43	-0.35
Des	0.675	-0.18	0.569	0.758		-0.38	-0.32
Skewness	0.052	0.236	-0.54	-0.43	-0.38		0.745
Kurtosis	-0.05	-0.06	-0.23	-0.35	-0.32	0.745	

Cuadro 4.4: Matriz de correlación para características de contraste del fondo

Características contraste relativo:
El cuadro 4.5 muestra la matriz de correlación para las características contraste relativo. Las características de contraste relativo y contraste proporcional muestran correlación al alcanzar un nivel de 0.983 y se procede a eliminar una de ellas.

	Absoluto	Relativo	Proporcional
Absoluto		0.841	0.827
Relativo	0.841		0.983
Proporcional	0.827	0.983	

Cuadro 4.5: Matriz de correlación para características de contraste relativo

Entre el contraste relativo y proporcional, este último tiene una ganancia de información mayor, por lo que se elimina el **contraste relativo**

Características de forma 1:
El cuadro 4.6 presenta los niveles de correlación para el primer grupo de características de forma, siendo estas: área, área convexa, área de fondo, área rellena, perímetro y diámetro mayor. Se determinó que existe correlación entre 4 pares de variables: área y área convexa, con nivel de 0.998; área y área rellena, con un nivel de 1; área rellena y área convexa con un nivel de 0.998; y diámetro y perímetro con una correlación de 0.974.

Entre el área, área convexa y área rellena, el área tiene una mayor ganancia de información y esta correlacionada con las dos ultimas, por lo que tanto **área convexa** como **área rellena** son eliminadas. Por otro lado, el **diámetro mayor** tiene una menor ganancia de información que el perímetro, por lo que también se elimina este diámetro.

Características de forma 2:
El cuadro 4.7 muestra los niveles de correlación para el grupo 2 de características de forma, siendo estas: diámetro menor, orientación, excentricidad, número de Euler, diámetro equivalente y solidez. Las características de diámetro equivalente y diámetro menor guardan correlación al poseer un nivel de 0.977 y una de ellas puede ser eliminada.

	Área	Área convexa	Á. fondo	Á. rellena	Perímetro	D. mayor
Área		0.998	0.889	1	0.930	0.951
Á. convexa	0.998		0.908	0.998	0.930	0.949
Á. fondo	0.889	0.908		0.889	0.783	0.807
Á. rellena	1	0.998	0.889		0.930	0.951
Perímetro	0.930	0.930	0.783	0.930		0.974
D. mayor	0.951	0.949	0.807	0.951	0.974	

Cuadro 4.6: Matriz de correlación para características de Forma 1

	D. menor	Orientación	Excentricidad	Euler	D. equivalente	Solidez
D. menor		-0.19	0.147	-0.05	0.977	0.266
Orientación	-0.19		0.057	-0.03	-0.02	-0.24
Excentricidad	0.147	0.057		-0.11	0.177	0.016
Euler	-0.05	-0.03	-0.11		-0.05	0.275
D. equivalente	0.977	-0.02	0.177	-0.05		0.245
Solidez	0.266	-0.24	0.016	0.275	0.245	

Cuadro 4.7: Matriz de correlación para características de Forma 2

De este segundo grupo de características de forma se elimina al **diámetro equivalente** debido a que tiene una menor ganancia de información que el diámetro menor.

Características de forma 3:

El cuadro 4.8 presenta los niveles de correlación para el tercer grupo de características de forma, siendo estas: redondez, relación de aspecto, elongación, compacidad 1, compacidad 2 y compacidad 3. Se determinó que existe correlación entre los pares: redondez y elongación, con un nivel de 0.97; redondez y compacidad 1, con 0.997, relación de aspecto y elongación con un nivel de -0.97; y compacidad 2 y compacidad 3 con un valor igual a la unidad.

	Redondez	Aspecto	Elongación	Compa. 1	Compa. 2	Compa. 3
Redondez		-0.94	0.970	0.997	-0.41	-0.41
Aspecto	-0.94		-0.97	-0.95	0.236	0.236
Elongación	0.970	-0.97		0.966	-0.24	-0.24
Compa. 1	0.997	-0.95	0.966		-0.42	-0.42
Compa. 2	-0.41	0.236	-0.24	-0.42		1
Compa. 3	-0.41	0.236	-0.24	-0.42	1	

Cuadro 4.8: Matriz de correlación para características de Forma 3

De este tercer y último grupo de característica de forma de la masa se elimina inicialmente la **redondez** y **elongación** por tener una menor ganancia de información que compacidad y relación de aspecto. En cuanto a compacidad 2 y compacidad 3, ambos tiene la misma ganancia de info, por lo que se elimina **compacidad 3**.

Momentos de secuencia de contorno:

El cuadro 4.9 muestra los niveles de correlación para los momentos de secuencia de contorno. Se determinó que existe correlación entre el MSC 1 y MSC3 al alcanzar un nivel de correlación de 0.993. También se encontró correlación entre el MSC 2 y MSC 4 con un nivel de 0.975. Se procedió a eliminar una característica de cada par de acuerdo a su ganancia de información.

	MSC 1	MSC 2	MSC 3	MSC 4	Prom radios	Des radios
MSC 1		0.855	0.993	0.917	0.018	0.609
MSC 2	0.855		0.875	0.975	-0.00	0.503
MSC 3	0.993	0.875		0.936	-0.00	0.583
MSC 4	0.917	0.975	0.936		-0.02	0.526
Prom radios	0.018	-0.00	-0.00	-0.02		0.687
Des radios	0.609	0.503	0.583	0.526	0.687	

Cuadro 4.9: Matriz de correlación para los momentos de secuencia de contorno MSC

El análisis de ganancia de información dice que el momento de secuencia de contorno 3 y 4 aportan más información que los momentos 1 y 2, respectivamente. Se eliminan **momentos de secuencia de contorno 1** y **momentos de secuencia de contorno 2**.

Momentos invariantes:

El cuadro 4.10 presenta los niveles de correlación para los primeros 7 momentos invariantes o momentos de Hu. El nivel de correlación más alto lo presentaron el momento invariante 4 (Hu 4) y momento invariante 6 (Hu 6) al obtener un nivel superior al 0.97 y por lo que se siguió a eliminar uno de ellos.

	Hu 1	Hu 2	Hu 3	Hu 4	Hu 5	Hu 6	Hu 7
Hu 1		0.902	0.768	0.732	0.498	0.647	0.511
Hu 2	0.902		0.594	0.641	0.429	0.596	0.409
Hu 3	0.768	0.594		0.928	0.757	0.848	0.839
Hu 4	0.732	0.641	0.928		0.862	0.972	0.859
Hu 5	0.498	0.429	0.757	0.862		0.926	0.947
Hu 6	0.647	0.596	0.848	0.972	0.926		0.871
Hu 7	0.511	0.409	0.839	0.859	0.947	0.871	

Cuadro 4.10: Matriz de correlación para los momentos invariantes

En el último paso del análisis de ganancia de información se elimina el **momento invariante 6** por aportar menos información que el momento invariante 4.

	Característica	Ganancia Promedio	Eliminar
1	Nivel de gris máximo	0.132	
9	Nivel de gris mínimo	0.135	
22	Perímetro	0.127	
24	Diámetro menor	0.122	
18	Área	0.115	
21	Área rellena	0.115	si
10	Mediana del nivel de gris del fondo	0.107	
31	Factor de forma	0.108	
5	Desviación estándar del nivel de gris	0.103	
28	Diametro equivalente	0.093	si
42	Promedio de radios	0.092	
19	Área convexa	0.091	si
23	Diámetro mayor	0.091	si
12	Desviación estándar del nivel de gris del fondo	0.09	
30	Alcance	0.087	
11	Promedio del nivel de gris del fondo	0.078	
40	Momento de secuencia de contorno 3	0.07	
44	Momento invariante 1	0.068	
3	Mediana del nivel de gris	0.068	
38	Momento de secuencia de contorno 1	0.065	si
29	Solidez	0.062	
6	Skewness	0.059	
2	Nivel de gris mínimo	0.055	
17	Contraste proporcional	0.057	
16	Contraste relativo	0.057	si
8	Nivel de gris máximo del fondo	0.054	
36	Compacidad 2	0.053	
4	Promedio del nivel de gris	0.052	si
37	Compacidad 3	0.053	si
26	Excentricidad	0.05	
41	Momento de secuencia de contorno 4	0.043	
25	Orientación	0.043	
43	Desviación estándar de radios	0.04	
47	Momento invariante 4	0.039	
39	Momento de secuencia de contorno 2	0.036	si
27	Número de Euler	0.031	
20	Área del fondo	0.031	
35	Compacidad 1	0.029	
46	Momento invariante 3	0.028	
13	Skewness del fondo	0.027	
32	Redondez	0.027	si

Cuadro 4.11: Ganancia de información

Característica	Ganancia Promedio	Eliminar
15 Contraste absoluto	0.027	
45 Momento invariante 2	0.025	
49 Momento invariante 6	0.025	si
33 Relación de aspecto	0.023	
7 Kurtosis	0.022	
34 Elongación	0.019	si
50 Momento invariante 7	0.017	
14 Kurtosis del fondo	0.012	
48 Momento invariante 5	0.005	

Cuadro 4.11: Ganancia de información

El cuadro 4.12 resume las características descartadas como resultado del análisis de correlación. En total se eliminaron 12 variables como resultado de este análisis, reduciendo el conjunto de 50 a 38 características.

Característica
Características de contraste de la señal
4 Promedio del nivel de gris
Características de contraste relativo
16 Contraste relativo
Características de forma
19 Área convexa
21 Área rellena
23 Diámetro mayor
28 Diametro equivalente
32 Redondez
34 Elongación
37 Compacidad 3
Momentos de secuencia de contorno
38 Momento de secuencia de contorno 1
39 Momento de secuencia de contorno 2
Momentos invariantes
49 Momento invariante 6

Cuadro 4.12: Resumen de características eliminadas en el análisis de correlación

Las características que eliminan en este análisis de correlación son nuevamente consideradas dentro de la búsqueda por medio del algoritmo genético, con el objetivo de reafirmar la validez del análisis. Se realizan dos búsquedas, una excluyendo este conjunto y una con la totalidad de características. La siguiente sección describe los

resultados obtenidos en la búsqueda evolutiva del espacio formado por subconjutnos de características.

4.4.2. Metodología de la envoltura (Wrapper) para evaluar subconjuntos de características

Una vez calculados los valores de las características que describen a masas, se procede a determinar un subconjunto que optimice el desempeño de los algoritmos: Redes Neuronales (NN), Máquinas de Vector de Soport (SVM), Análisis de Discriminante Lineal (LDA) y Regresión Logística (LR). Para ello se utilizan algoritmos genéticos, descritos en la sección 2.3.5, que brindan una solución subóptima sin la necesidad de explorar todo el conjunto de soluciones de las combinaciones posibles. Al final de la ejecución de dicho algoritmo la mejor solución obtenida representa el subconjunto de características seleccionado y corresponde al algoritmo clasificador con el mejor desempeño desempeño dado un conjunto de entradas.

Red Neuronal (NN)

Para la experimentación se corrieron diferentes algoritmos genéticos, con un tamaño de cromosoma igual a 50, es decir, el número de características calculadas; una población de 70 individuos y cada uno se dejó evolucionar durante 100 generaciones. Los parámetros que se variaron para los diferentes experimentos fueron la probabilidad de mutación p_m, con valores entre 0.1 y 0.4, y la probabilidad de cruce p_c, con valores entre 0.7 y 0.9. Como método de selección se utilizó el torneo de tamaño 2. También se varió las épocas de entrenamiento de la red neuronal, donde inicialmente se consideraban 250 y luego 50 épocas.

El cuadro 4.13 muestra la matriz de confusión obtenida con el mejor individuo encontrado en la búsqueda mediante algoritmos genéticos. De los 85 ejemplos de masas con diagnóstico de cáncer, el individuo seleccionado por el algoritmo clasifica correctamente a 70. En cuanto a las masas benignas detectadas en las mamografías, el algoritmo la red construida con las características representadas en el mejor individuo del algoritmo genético clasifica correctamente 68 de 95.

		Patología	
		Positivo	**Negativo**
Resultado de la prueba	**Positivo**	70	27
	Negativo	15	68

Cuadro 4.13: Matriz de confusión: Red Neuronal

El cuadro 4.14 muestra los indicadores de rendimiento calculados para red neuronal que considera las características codificadas en el mejor individuo encontrado por el

algoritmo genético. La red obtuvo un valor de 82.35 % de sensibilidad mientras que un 70.65 % de especificidad.

Éxito	Sensibilidad	Especificidad	Error
76.27	82.35	70.65	23.73

Cuadro 4.14: Rendimiento: Red neuronal

La características seleccionadas para el algoritmo de red neuronal se muestran en el cuadro 4.15.

	Característica
	Características de contraste de la señal
1	Nivel de gris máximo
2	Nivel de gris mínimo
4	Promedio del nivel de gris
5	Desviación estándar del nivel de gris
6	Asimetría del nivel de gris (Skewness)
	Características de contraste del fondo
8	Nivel de gris máximo
9	Nivel de gris mínimo
12	Desviación estándar del nivel de gris
	Características de contraste relativo
16	Contraste relativo
	Características de forma
18	Área
19	Área convexa
20	Área del fondo
24	Diámetro menor
25	Orientación
27	Número de Euler
28	Diámetro circular equivalente
33	Relación de aspecto
35	Compacidad 1
37	Compacidad 3
	Momentos de secuencia de contorno
38	Momento de secuencia de contorno 1
42	Promedio de radios
	Momentos invariantes
44	Momento invariante 1
47	Momento invariante 4
48	Momento invariante 5

Cuadro 4.15: Características seleccionadas para red neuronal

Como resultado de la selección de características, se redujo de 50 a 24 las características a considerar para el clasificador de red neuronal. Entre las características sobrevivientes al proceso de selección se encuentra al menos una de cada grupo de características, siendo las características de forma las que mayor número aporta al total.

Los resultados recién mostrados corresponden a la búsqueda evolutiva realizada con la totalidad de características, es decir, considerando a las sugeridas para descartar en el análisis de correlación. Como resultado de este proceso se incluye dentro del modelo las características de *promedio del nivel de gris, contraste relativo, área convexa, diámetro circular equivalente, compacidad 3* y *momento de secuencia de contorno 1.*

Máquina de vector de soporte (SVM)

Para la experimentación se corrieron diferentes algoritmos genéticos, con un tamaño de cromosoma igual a 50, es decir, el número de características calculadas; una población de 500 individuos y cada uno se dejó evolucionar durante 50 generaciones. Se definió una p_m igual a 0.3 y p_c de 0.9. La selección se realizó mediante torneo binario con parejas aleatorias.

El cuadro 4.16 muestra los resultados obtenido en la selección de características con la máquina de vectores de soporte. El clasificador clasifico correctamente 76 del total de caso positivos (masas malignas) y 77 de los negativos (masas benignas).

		Patología	
		Positivo	**Negativo**
Resultado de la prueba	**Positivo**	76	18
	Negativo	9	77

Cuadro 4.16: Matriz de confusión: Máquina de vector de soporte

Las medidas de desempeño de la máquina de vectores de soporte con mejor aptitud al resto se muestran en el cuadro 4.17. El SVM se encuentra levemente por debajo del 90 % de sensibilidad y arriba del 80 % de especificidad con una medida de éxito de 85.0 %.

Éxito	**Sensibilidad**	**Especificidad**	**Error**
85.00	89.41	81.05	15.00

Cuadro 4.17: Rendimiento: Máquina de vector de soporte

Se seleccionaron un total de 22 características las cuales se muestran en el cuadro 4.18:

	Característica
	Características de contraste de la señal
1	Nivel de gris máximo
3	Mediana del nivel de gris
5	Desviación estándar del nivel de gris
6	Asimetría del nivel de gris (Skewness)
7	Kurtosis del nivel de gris
	Características de contraste del fondo
8	Nivel de gris máximo
10	Mediana del nivel de gris
11	Promedio del nivel de gris
13	Asimetría del nivel de gris (Skewness)
	Características de contraste relativo
15	Contraste absoluto
	Características de forma
18	Área
24	Diámetro menor
25	Orientación
26	Excentricidad
36	Compacidad 2
	Momentos de secuencia de contorno
40	Momento de secuencia de contorno 3
41	Momento de secuencia de contorno 4
42	Promedio de radios
	Momentos invariantes
44	Momento invariante 1
45	Momento invariante 2
46	Momento invariante 3
47	Momento invariante 4

Cuadro 4.18: Características seleccionadas para máquina de vector de soporte

Se corrieron los algoritmos con cromosoma de tamaño 50, es decir, considerando la totalidad de características y posteriormente se realizó la experimentación eliminando inicialmente las 12 características sugeridas para ser eliminadas. En el caso de las máquinas de vectores de soporte los mejores resultados provienen de la combinación del análisis discriminante y la selección mediante algoritmos genéticos.

Del grupo de características de contraste de la masa (o señal) se seleccionaron 5 características mientras que de las características de contraste del fondo 4. Se seleccionó una característica de contraste relativo, 4 características de forma, 3 del grupo de momentos de secuencia de contorno y 4 de los 7 momentos invariantes de Hu.

Análisis discriminantes (LDA)

Para la experimentación se corrieron diferentes algoritmos genéticos, con dos tamaños de cromosoma: el primero igual a 50, es decir, el número de características calculadas; y el segundo con un tamaño de 38, correspondiente al total de características menos las eliminadas en el análisis de correlación. La población usada fue de 500 individuos y cada uno se dejó evolucionar durante 100 generaciones. Los parámetros que se variaron para los diferentes experimentos fueron la probabilidad de mutación p_m, con valores entre 0.1 y 0.4, y la probabilidad de cruce p_c, con valores entre 0.7 y 0.9. Como método de selección se utilizó el torneo de tamaño 2.

La matriz de confusión del individuo con mejor desempeño se muestra en el cuadro 4.19. De entre las masas con diagnóstico maligno, se clasificó correctamente a 65, lo que representa que 20 fueron clasificadas de manera errónea. Del total de 95 casos negativos (masas benignas), se clasificó correctamente a 66 de los casos.

		Patología	
		Positivo	**Negativo**
Resultado de la prueba	**Positivo**	65	29
	Negativo	20	66

Cuadro 4.19: Matriz de confusión: Análisis discriminante

El clasificador LDA obtuvo un valor de sensibilidad de 76.47 % y uno de especificidad de 69.47 %. El error del clasificador supera el 25 %, lo que representa un valor considerablemente alto. Los resultados completo se muestran en el cuadro 4.20.

Éxito	**Sensibilidad**	**Especificidad**	**Error**
72.78	76.47	69.47	27.22

Cuadro 4.20: Rendimiento: Análisis discriminante

A continuación se presentan, en el cuadro 4.21, las 14 características seleccionadas en el proceso:

	Característica
	Características de contraste de la señal
1	Nivel de gris máximo
2	Nivel de gris mínimo
3	Mediana del nivel de gris
5	Desviación estándar del nivel de gris
7	Kurtosis del nivel de gris

Cuadro 4.21: Características seleccionadas para análisis discriminante

	Característica
	Características de contraste del fondo
8	Nivel de gris máximo
11	Promedio del nivel de gris
12	Desviación estándar del nivel de gris
	Características de forma
36	Compacidad 2
	Momentos de secuencia de contorno
41	Momento de secuencia de contorno 4
42	Promedio de radios
43	Desviación estándar de radios
	Momentos invariantes
44	Momento invariante 1
50	Momento invariante 7

Cuadro 4.21: Características seleccionadas para análisis discriminante

Como resultado de la búsqueda, el algoritmo genético excluyó la totalidad de atributos del grupo de características de contraste relativo y redujo a una el grupo de características de forma de la masa. De un total de 50 el algoritmo determinó que el individuo con mejor rendimiento consideraba únicamente 14 características.

Los resultados obtenidos para análisis discriminante (LDA), se obtuvieron de la aplicación secuencial del análisis de correlación y la selección mediante algoritmo genético.

Regresión logística (LR)

Para la experimentación se corrieron diferentes algoritmos genéticos, con un tamaño de cromosoma igual a 50, es decir, el número de características calculadas; una población de 500 individuos y cada uno se dejó evolucionar durante 50 generaciones. Se definió una p_m igual a 0.3 y p_c de 0.9. La selección se realizó mediante torneo binario con parejas aleatorias.

Como resultado del algoritmo genético, el mejor individuo encontrado en la búsqueda, clasifico correctamente 65 y 74, ejemplos positivos (masa maligna) y negativos (masa beninga), respectivamente. La matriz de confusión completa se muestra en el cuadro 4.22.

El cuadro 4.23 lista los valores de desempeño obtenidos con el mejor encontrado en la búsqueda mediante algoritmos genéticos y el modelo de la envoltura para la regresión logística LR. En este caso el mejor individuo obtuvo un valor de especificidad de 77.89 %, levemente superior a su sensibilidad de 76.47 % para obtener un 77.22 de éxito.

		Patología	
		Positivo	**Negativo**
Resultado de la prueba	**Positivo**	65	21
	Negativo	20	74

Cuadro 4.22: Matriz de confusión: Regresión logística

Éxito	**Sensibilidad**	**Especificidad**	**Error**
77.22	76.47	77.89	22.78

Cuadro 4.23: Rendimiento: Regresión Logística

De las 50 características extraídas inicialmente se seleccionaron 28, las cuales se listan en el cuadro 4.24.

	Característica
	Características de contraste de la señal
1	Nivel de gris máximo
2	Nivel de gris mínimo
3	Mediana del nivel de gris
4	Promedio del nivel de gris
5	Desviación estándar del nivel de gris
7	Kurtosis del nivel de gris
	Características de contraste del fondo
9	Nivel de gris mínimo
10	Mediana del nivel de gris
11	Promedio del nivel de gris
12	Desviación estándar del nivel de gris
	Características de contraste relativo
15	Contraste absoluto
	Características de forma
23	Diámetro mayor
24	Diámetro menor
25	Orientación
26	Excentricidad
28	Diámetro circular equivalente
29	Solidez
30	Alcance
32	Redondez
35	Compacidad 1
	Momentos de secuencia de contorno
38	Momento de secuencia de contorno 1
39	Momento de secuencia de contorno 2

Cuadro 4.24: Características seleccionadas para regresión logística

	Característica
41	Momento de secuencia de contorno 4
42	Promedio de radios
	Momentos invariantes
46	Momento invariante 3
47	Momento invariante 4
49	Momento invariante 6
50	Momento invariante 7

Cuadro 4.24: Características seleccionadas para regresión logística

Los resultados reportados para regresión logística se obtuvieron mediante la búsqueda en el espacio completo de características.

Rendimiento algoritmos de clasificación en la selección de características

La figura 4.1 ubica a los mejores individuos para cada uno de los cuatro algoritmos de clasificación considerados en el espacio ROC. El algoritmo que mejor desempeño mostró fue el SVM y el de peor LDA. Los algoritmos NN y LR tienen en general un desempeño similar, aunque la red presenta una brecha mayor entre sus medidas de sensibilidad y especifidad.

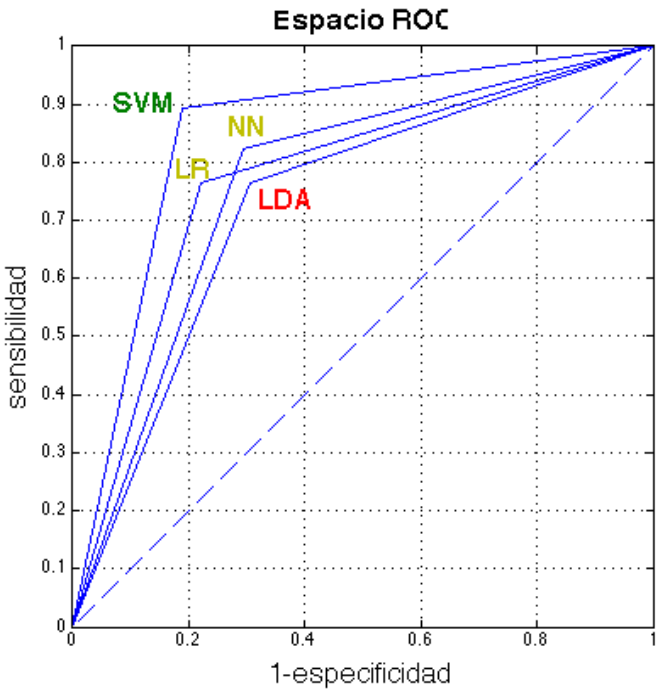


Figura 4.1: Rendimiento algoritmos de clasificación en la selección de características

4.5. Clasificación

Una vez seleccionadas las características, se procede con la última fase del modelo de solución planteado en la presente investigación. En este punto se realiza un conjunto de experimentos para determinar el desempeño promedio de clasificación en función de las características de entrada y seleccionar un clasificador con menor error, como una segunda validación de las redes obtenidas en la fase de selección de características.

4.5.1. Red Neuronal (NN)

Se construyó un total 100 redes neuronales, donde los ejemplos utilizados en el aprendizaje consideraban únicamente las características que se listan en el cuadro 4.15. Los mejores 10 resultados de la validación cruzada de 10 conjuntos se listan en el cuadro 4.25. Estos resultados fueron ordenados de acuerdo al éxito de los clasificadores. De la experimentación se puede afirmar que la red neuronal tiene en promedio un 77.18 % de sensibilidad y 70.74 % de especididad, lo que equivale un error estimado al clasificar de 26.22 %.

El proceso de clasificación con mejores resultados obtuvo un rendimiento levemente inferior al obtenido por el mejor individuo encontrado en la selección de características, con una diferencia marcada en el la especificidad, y por lo tanto en el el porcentaje de éxito.

Ranking	**Éxito**	**Sensibilidad**	**Especificidad**	**Error**
1	0.7611	0.8118	0.7158	0.2389
2	0.7500	0.7765	0.7263	0.2500
3	0.7500	0.7412	0.7579	0.2500
4	0.7444	0.8000	0.6947	0.2556
5	0.7389	0.8000	0.6842	0.2611
6	0.7333	0.7882	0.6842	0.2667
7	0.7278	0.7294	0.7263	0.2722
8	0.7278	0.7294	0.7263	0.2722
9	0.7222	0.7529	0.6947	0.2778
10	0.7222	0.7882	0.6632	0.2778
Media	0.7378	0.7718	0.7074	0.2622
Des. Estàndar	0.0133	0.0310	0.0280	0.0133

Cuadro 4.25: Clasificación: Red Neuronal

Los valores de desviación estándar de sensibilidad y especificidad son considerablemente altos comparados con los obtenidos en el éxito y error. Esto muestra las variaciones que se perciben en el rendimiento del clasificador.

4.5.2. Máquina de vector de soporte (SVM)

Para validar los resultados obtenidos en la etapa de selección de características se construyeron 100 modelos de máquinas de vector de soporte, donde los ejemplos considerados para el aprendizaje están formados por las características obtenidas en la reducción de dimensiones y que se detallan en el cuadro 4.18.

El cuadro 4.26 muestra los mejores 10 resultados obtenidos para la validación cruzada de 10 conjuntos y el clasificador SVM. El valor promedio de sensibilidad se encuentra por debajo del 89 %, que corresponde a la sensibilidad del mejor encontrado en la búsqueda mediante algoritmo genético para la selección de características. En promedio el clasificador SVM presento una sensibilidad de 86.71 %, una especificidad de 78.95 % y un porcentaje de éxito de 82.61 %.

Ranking	Éxito	Sensibilidad	Especificidad	Error
1	0.8500	0.8941	0.8105	0.1500
2	0.8444	0.8824	0.8105	0.1556
3	0.8389	0.8824	0.8000	0.1611
4	0.8333	0.8706	0.8000	0.1667
5	0.8222	0.8706	0.7789	0.1778
6	0.8222	0.8588	0.7895	0.1778
7	0.8222	0.8588	0.7895	0.1778
8	0.8167	0.8588	0.7789	0.1833
9	0.8111	0.8471	0.7789	0.1889
10	0.8000	0.8471	0.7579	0.2000
Media	0.8261	0.8671	0.7895	0.1739
Des. Estàndar	0.0155	0.0157	0.0165	0.0155

Cuadro 4.26: Clasificación: Máquina de vector de soporte

4.5.3. Análisis discriminantes (LDA)

Al igual que con los dos clasificadores previos, se construyó un total 100 modelos de análisis de discriminante cuyos ejemplos de entrenamiento consideraban únicamente las características que se listan en el cuadro 4.21. Los mejores resultados se listan de forma ordenada en el cuadro 4.27.

Los resultados obtenidos con el análisis discriminantes en la etapa de clasificación se encuentran bastante por debajo de los obtenidos en la etapa de selección de características. El valor de sensibilidad es muy malo, al grado de ser comparable con la predicción del resultado al lanzar una moneda al aire. El promedio de los mejores 10 resultados de validación para el análisis discriminante se resume en un promedio de 62 % de sensibilidad, 72.42 % de especificidad, 67.50 % de y 32.50 % de error.

La desviación estándar de las cuatro métricas de desempeño muestra como el rendimiento del clasificador es constantemente bajo.

Ranking	Éxito	Sensibilidad	Especificidad	Error
1	0.6889	0.6353	0.7368	0.3111
2	0.6833	0.6235	0.7368	0.3167
3	0.6833	0.6353	0.7263	0.3167
4	0.6778	0.6353	0.7158	0.3222
5	0.6778	0.6235	0.7263	0.3222
6	0.6778	0.6235	0.7263	0.3222
7	0.6667	0.6000	0.7263	0.3333
8	0.6667	0.6118	0.7158	0.3333
9	0.6667	0.6235	0.7053	0.3333
10	0.6611	0.5882	0.7263	0.3389
Media	0.6750	0.6200	0.7242	0.3250
Des. Estàndar	0.0092	0.0157	0.0097	0.0092

Cuadro 4.27: Clasificación: Análisis discriminante

4.5.4. Regresión logística (LR)

Para validar los resultados obtenidos en la etapa de selección de características se construyeron 100 modelos de regresión logística, considerando como entrada las características obtenidas en la reducción de dimensiones y que se detallan en el cuadro 4.24.

El modelo de regresión logística fue bastante estable a lo largo de los 100 experimentos de clasificación. Los resultados promedio de clasificación son inferiores a los obtenidos en la etapa de selección de características.

El modelo de regresión logística obtuvo en promedio un 73 % para éxito, sensibilidad y especificidad.

Ranking	Éxito	Sensibilidad	Especificidad	Error
1	0.7389	0.7529	0.7263	0.2611
2	0.7389	0.7412	0.7368	0.2611
3	0.7333	0.7412	0.7263	0.2667
4	0.7333	0.7412	0.7263	0.2667
5	0.7278	0.7412	0.7158	0.2722
6	0.7278	0.7412	0.7158	0.2722
7	0.7444	0.7294	0.7579	0.2556
8	0.7444	0.7294	0.7579	0.2556
9	0.7444	0.7294	0.7579	0.2556
10	0.7389	0.7294	0.7474	0.2611
Media	0.7372	0.7376	0.7368	0.2628
Des. Estándar	0.0064	0.0079	0.0172	0.0064

Cuadro 4.28: Clasificación: Regresión logística

Rendimiento algoritmos en la etapa de clasificación

El cuadro 4.29 resume los resultados de la etapa de clasificación para los cuatro algoritmos considerados en la presente investigación:

Clasificador	Éxito	Sensibilidad	Especificidad	Error
Red neuronal NN	0.7378	0.7718	0.7074	0.2622
Máquina de vector de soporte SVM	0.8261	0.8671	0.7895	0.1739
Análisis discriminante LDA	0.6750	0.6200	0.7242	0.3250
Regresión Logística LR	0.7372	0.7376	0.7368	0.2628

Cuadro 4.29: Resumen de clasificación

La figura 4.2 ubica los resultados promedio obtenidos para cada uno de los cuatro algoritmos de clasificación considerados en el espacio ROC. El algoritmo que mejor desempeño mostró fue el SVM y el de peor LDA. Los algoritmos NN y LR tienen en general un desempeño similar, aunque la red presenta una brecha mayor entre sus medidas de sensibilidad y especificidad. Esta tendencia es la misma observada en los mejores individuos obtenidos en la etapa de selección de características.

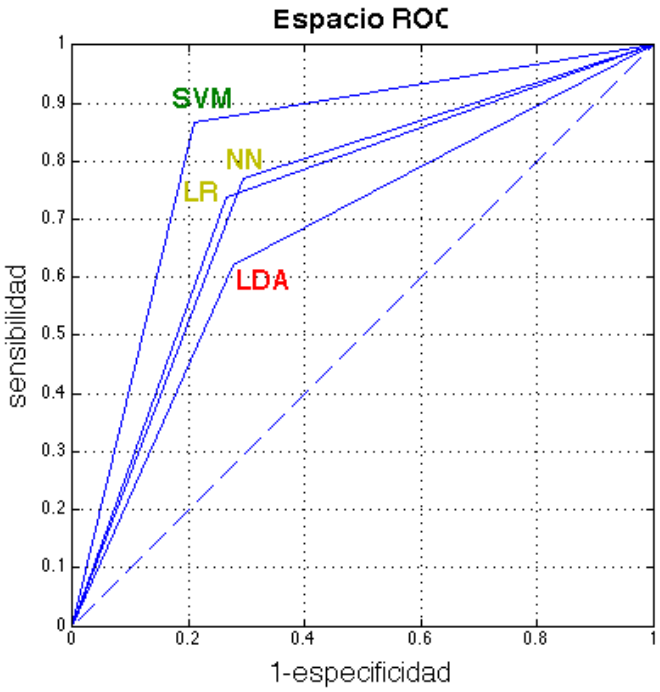


Figura 4.2: Rendimiento algoritmos en la etapa de clasificación

El rendimiento de los cuatro algoritmos de clasificación es menor en la etapa de clasificación al obtenido por el mejor individuo en la selección de características. El porcentaje de éxito promedio de cada uno de los clasificadores en la etapa de clasificación se

encuentra aproximadamente un 4 % por debajo de los resultados obtenidos en selección de características.

Situación similar a la ocurrida con el porcentaje de éxito ocurre con la sensibilidad de los algoritmos de red neurnoal NN, máquina de vector de soporte SVM y regresión logística LR, donde se presenta una reducción del valor de dicha métrica de entre 3 y 5 %. Mismo fenómeno se observa en la especificidad del SVM y LR, con una baja en su valor de aproximadamente 3 %. En el caso de red neuronal, la especificidad obtenida en la etapa de clasificación es valor similar al de la etapa de selección de características.

El modelo del clasificador SVM es el que mejor desempeño obtuvo en la etapa de selección de características y en la clasificación. Su porcentaje de error promedio es de un 17.30 % y su sensibilidad, la más alta de los cuatro clasificadores es en promedio un 86.71 %. La especificidad de este clasificador tiene un valores levemente inferior al 80 %.

El modelo de regresión logística fue el más estable a lo largo de los experimentos de clasificación al grado que la variación en su desempeño es considerablemente menor si se compara con como cambiaron las métricas de rendimiento de los otros tres algoritmos de clasificación. Esto se puede observar en los valores de desviación estándar de los indicadores los cuales presentan valor por debajo del 1 porcentual.

El algoritmo con le menor desempeño fue el algoritmo cuyo conjunto de características tiene el menor tamaño, siendo este el análisis discriminante LDA. Si bien el porcentaje de éxito este clasificador disminuyó un 5 %, el resultado con el cambio más significativo fue la sensibilidad, donde el mejor individuo de la selección de características estaba arriba del 75 % y en la etapa de clasificación bajó hasta un 62 %, siendo éste último el peor de los resultados obtenidos por alguno de los clasificadores.

El rendimiento del clasificador SVM puede ser comparado con investigaciones que han obtenido buenos resultados. Trabajos previos indican un porcentaje de éxito para clasificación de masas tumorales en maligno y benigno aproximado a 84 % [64] por un 82 % obtenido en esta investigación. Otras publicaciones indican un valor de sensibilidad aproximado a 81 % [49] versus un promedio de 86.71 % obtenido en la presente investigación con el clasificador SVM. Estos datos indican que los resultados obtenidos con el clasificador SVM se encuentran dentro de rangos adecuados al compararlos con otros modelos de solución.

En el caso de los clasificadores NN y LDA, investigaciones previas reportan porcentajes de éxito de 69.23 % para NN y 63.46 % para LDA [20]. En la presente investigación se determinó un conjunto de características para NN que en promedio entrega un 73.78 % de éxito al clasificar. El modelo para LDA determinado en esta investigación tiene un rendimiento similar al reportado en [20], con un 62 % de éxito. El contraste con respecto a esta investigación radica en la especificidad obtenida en el modelo de esta investigación donde NN obtuvo un 70 % y LDA un 72 %.

En el estudio presentado [5] se reporta una especificidad de 77.9 % para regresión logística contra un 73.68 % obtenido en la presente investigación. El mismo estudio muestra una sensibilidad de 80.50 % por un 73.76 % promedio obtenido en esta investigación.

4.6. Resumen

En el presente capítulo del documento se describen los experimentos realizados durante la aplicación del modelo de solución y se muestran los resultados obtenidos en cada uno de ellos.

En primer lugar, se realiza un análisis general de la base de datos DDSM, que es utilizada en la presente investigación. Se considera un total de 180 mamografías, 99 cráneo-caudal CC y 81 medio-lateral oblicua MLO, provenientes de 119 casos donde se detectó una masa delimitada por único borde, de las cuales 95 tienen un diagnóstico de benigno y 85 maligno. Se realiza un análisis visual para determinar un nivel de gris que dé la forma aproximada de la masa dentro de la zona con anomalía que indica la base de datos.

Como siguiente fase del modelo de solución, y con base al análisis de la base de datos DDSM, para cada una de las imágenes con masas seleccionadas se extrae un total de 50 características y se presenta el resumen de los valores obtenidos, indicando estadísticos como valor mínimo, valor máximo, media y desviación estándar.

Una vez construida la base de datos de ejemplos para el aprendizaje se procede a la etapa de reducción de dimensiones. La primera parte del modelo consiste en un análisis de correlación de las características obtenidas. Se presentan las diferentes matrices de correlación para cada grupo de características, resaltando los pares de características con un alto nivel de correlación. Posteriormente se presenta el análisis de ganancia de información, donde las características son ordenadas de acuerdo a su valor de ganancia de información, lo que ofrece la posibilidad de eliminar la característica de cada pareja correlacionada con menor aporte de información. Como resultado de este análisis se determino correlación estadística entre 12 pares de variables del modelo, lo que permitió reducir el conjunto de características de 50 a 38.

La siguiente segunda parte de la etapa de reducción de dimensiones es la etapa de selección de características la cual se realizó mediante el modelo de envoltura (*wrapper*) que utilizó como algoritmo de búsqueda un algoritmo genético. El proceso se llevo a cabo con cuatro algoritmos de clasificación: redes neuronales NN, máquinas de vector de soporte SVM, análisis discriminante LDA y regresión logística LR. Para el clasificador de neuronal se determinó un conjunto de 24 características con una sensibilidad de 82.35 % y una especificidad de 70.65 %. Para la máquina de vector de soporte el conjunto de características se redujo hasta 22 las cuales mostraron un 89.41 % en sensibilidad y 81.05 % en especificidad. El análisis discriminante fue el algoritmo cuyo conjunto de características seleccionado es el de menor tamaño de entre los cuatro clasificadores, con una cardinalidad de 14 características, una sensibilidad de 76.47 % y una especificidad de 69.47 %. El último clasificador con el que se trabajó fue la regresión logística y para éste se determinó un conjunto de 28 características que proporcionan una sensibilidad del 76.47 % y una especificidad de 77.89 %.

La última etapa del modelo, la clasificación de masas, se desarrolló mediante la construcción de 100 modelos de cada clasificador, donde se consideraron únicamente las características seleccionadas en la etapa previa. En esta etapa la red neuronal obtuvo un porcentaje de éxito promedio de 73.78 %, la máquina de vector de soporte un 82.61 % y la regresión logística un 73.72 %. El análisis de discriminante fue el clasificador con el menor rendimiento con un 32.50 % de error.

Capítulo 5

Conclusiones y recomendaciones

El análisis de los resultados obtenidos durante el proceso del modelo de solución y de los experimentos realizados del presente trabajo, se describe en la sección de conclusiones y contribuciones en el presente capítulo. Adicionalmente se presentan algunas sugerencias para extender el área de investigación con trabajos futuros.

5.1. Conclusiones

La presente tesis describe el desarrollo de un modelo para la selección automática de características y clasificación de masas tumorales por medio de redes neuronales, máquinas de vector de soporte, análisis discriminante y regresión logística en mamografías digitales.

La investigación inicia con un estudio de la base de datos Digital Database for Screening Mammography (DDSM), para seleccionar los casos a considerar en la experimentación y validación del modelo propuesto. Una vez escogido el conjunto de casos a considerar se procede a utilizar la información que provee la DDSM para segmentar las masas de cada una de las mamografías, extrayendo así la región de interés ROI que contiene a la lesión. Luego se extraen características que describen su forma y textura, se lo analiza y se seleccionan un subgrupo representativo que permitan identificar si una masa localizada en una mamografía es diagnosticada como maligna (cancerosa) o benigna (no cancerosa). El conjunto de características extraído de cada masas se lleva a un proceso de reducción de dimensiones mediante análisis de correlación y ganancia de información; y selección de características mediante el modelo de envoltura con un algoritmo genético. La última etapa consiste en la validación de los conjuntos de características obtenidos para cada clasificador y obtener métricas de desempeño de los algoritmos.

El presente trabajo considera casos de la Digital Database for Screening Mammography (DDSM). Se efectuó un análisis de los casos contenidos en la base de datos DDSM de donde se decidió considerar un total de 180 mamografías provenientes de 119

casos donde se detectó una masa delimitada por único borde, de las cuales 95 tienen un diagnóstico de benigno y 85 maligno. 99 mamografías corresponden a la vista cráneo-caudal CC y de esas 51 tienen un diagnostico maligno. Las 81 mamografías restantes corresponde a la vista medio-lateral oblicua MLO, 34 contienen una masa con diagnóstico de cáncer. De las 85 masas diagnosticadas como malignas, 72 corresponden a masas con margen espiculado. En este caso se consideraron únicamente casos donde las masas detectadas tenías un nivel de apreciación superior o igual a 4.

Para los casos de masas seleccionados, se toma la mamografía, se pre-procesa mediante un filtro de mediana, con el objeto de eliminar el ruido contenido en ella. Posteriormente se hace uso de la información proporcionada por la base DDSM para delimitar la lesión en cuestión y posteriormente segmentar la región de interés que contiene a la masa. Una vez localizadas las señales, se procede a extraer características que la describen. Un total de 50 características son calculadas en esta fase, de las cuales 7 son de contraste de la señal, 7 son de contraste del fondo, 3 son de contraste relativo, 20 son de forma, 6 están relacionadas con momentos de la secuencia de contorno y 7 son los primeros momentos invariantes de Hu.

Dada la naturaleza de las características y con el objetivo de reducir el espacio de búsqueda de subconjuntos de características se llevó a cabo un análisis de correlación para determinar las parejas de características con un alto nivel de correlación, encontrándose 12 parejas de características altamente correlacionadas. Posteriormente se llevó a cabo un proceso de discretización de las características, para posteriormente obtener la ganancia de información de cada una de ellas. Con base en la ganancia de información, es posible eliminar de cada par de características correlacionadas aquella con la menor ganancia de información. Con esto el conjunto de características se puede reducir de 50 a 38. Este procedimiento depende de la relación entre características y la clase asociada a cada caso.

Para continuar con el proceso de reducción de características y obtener un subconjunto que sirva como entrada al clasificador a aplicar, se utiliza un modelo *wrapper* con algoritmos genéticos para explorar el conjunto de posibles soluciones. Para esta fase se utilizó validación cruzada de 10 conjuntos, como forma de verificar el conjunto de características seleccionadas que entrega el algoritmo genético. La métrica utilizada para medir el desempeño de los clasificadores generados por la validación cruzada es el porcentaje de éxito y la sensibilidad del clasificador. El proceso se lleva a cabo con cada uno de los cuatro clasificadores previamente listados.

Se determinó un conjunto 24 características para el clasificador de red neuronal. La reducción de características se llevo a cabo considerando las 50 características extraídas inicialmente como parte del espacio de búsqueda del algoritmo genético y se obtuvo como porcentaje de éxito al clasificar un 76.27 % y una sensibilidad de 82.35 %.

Para el algoritmo de máquina de vector de soporte se encontró que el conjunto con mayor poder para discriminar entre masas malignas y benignas tiene un tamaño de 22 características y es el resultado de la aplicación del análisis de correlación (con lo que se

reduce el espacio de 50 a 38 características) y la posterior búsqueda mediante algoritmo genético. El clasificador SVM obtuvo en la selección de características un 89.41 % de sensibilidad y un 81.05 % de especificidad, lo que implica un 15 % de error al clasificar, siendo el clasificador con los resultados más altos.

Como resultado del análisis de correlación y la selección de características, se determinó un conjunto de 14 características para el análisis discriminante lineal LDA, siendo éste el clasificador con el conjunto de características de menor tamaño y cuyo porcentaje de éxito en la etapa de selección fue de 72.78 %, con una sensibilidad de 76.47 % y especificidad de 69.47 %. Estos resultados fueron obtenidos maximizando la sensibilidad del modelo.

Para el modelo de regresión logística se obtuvo como resultado de la selección de características un conjunto de 28 características cuyas métricas de desempeño en la etapa de selección se resumen en un 77.22 % de éxito al clasificar con una sensibilidad de 76.47 % y una especificidad de 77.89 %. La regresión lineal es el modelo con el mayor número de características seleccionadas.

En la última etapa de la investigación se construyen 100 modelos de cada algoritmo clasificador y se obtienen las métricas de rendimiento promedio de *porcentaje de éxito*, *sensibilidad*, *especificidad* y *error* a través de la validación cruzada de 10 conjuntos para los 10 mejores resultados de cada clasificador.

De la etapa de clasificación se pudo determinar que el clasificador con mejor rendimiento fue la máquina de vector de soporte con un 82.61 % de porcentaje de éxito, 86.71 % de sensibilidad y una especificidad promedio de 78.95 %. Los clasificadores de red neuronal y regresión logística presentaron un rendimiento similar. La red neuronal obtuvo un 73.78 % de éxito, 77.18 % de sensibilidad y 70.74 % de especificidad con un 74 % aproximado tanto en éxito, como en sensibilidad y especificidad de la regresión logística. El modelo de análisis discriminante fue el que tuvo el desempeño más bajo en clasificación, con un porcentaje de error al clasificar de 32.50 %, un 62.00 % de sensibilidad y una especificidad 72.42 %.

El rendimiento del clasificador SVM se encuentra en el rango reportado en investigación similares, mientras que los algoritmos NN y LR presentaron resultados aceptables pero inferiores a los reportados en otros trabajos.

5.2. Contribuciones

El aporte principal del presente trabajo de investigación consiste en el desarrollo de un modelo de para la clasificación automática de masas tumorales en mamografías digitales.

De forma específica, las contribuciones logradas con el presente trabajo de investigación son:

- Analizar la base de datos DDSM para seleccionar un conjunto de casos de estudio

con mamografías digitales del tipo cráneo-caudal CC y medio-lateral oblicua MLO con masas tumorales identificadas.

- Construir una base de datos de masas tumorales en mamografías digitales mediante la extracción de características que representan de forma numérica el contenido visual de las lesiones.
- Obtener un ordenamiento de las características extraídas, tomando en consideración el indicador de ganancia de información.
- Determinar para cada uno de los clasificadores redes neuronales, máquinas de vector de soporte, análisis discriminante y regresión logística el conjunto, el conjunto de características que máximizan su desempeño en el porcentaje de éxito, sensibilidad y especificidad.
- Obtener métricas de desempeño de porcentaje de éxito, sensibilidad, especificidad y error para los modelos determinados para cada algoritmo de clasificación considerado.
- Brindar al radiólogo un sistema que apoye de manera automática el proceso de diagnóstico de lesiones de masas en mamografías digitales.

5.3. Trabajo Futuro

Entre los temas a considerar para ampliar la investigación se pueden mencionar:

- Aplicar las diferentes fases del modelo de solución a otras bases de datos, para analizar los diferentes parámetros calculados y los resultados en el presente documento.
- Implementar nuevas características que se puedan extraer desde cada mamografía.
- Ejecutar la selección de características con un mayor número de iteraciones para determinar si se puede encontrar un mejor número de características que generen una adecuada clasificación.
- Implementar la selección de características por medio de una búsqueda secuencial utilizando el ordenamiento de características de acuerdo a su ganancia de información.
- Utilizar otros clasificadores para comparar su desempeño y determinar la forma de mejorar el desempeño de los algoritmos considerados en esta investigación.
- Ampliar la experimentación para comparar entre el modelo de entrenamiento *off-line* e *in-line*.

Bibliografía

[1] Abeloff, M. D. *Cancer of the Breast*, 4â« ed. ed. Philadelphia, 2008, ch. 95.

[2] Alpaydin, E. *Introduction to machine learning*, 2nd ed. Adaptive computation and machine learning. MIT Press, Cambridge, Mass., 2010.

[3] American Cancer Society. Guía detallada: Cáncer de seno, 2011.

[4] Bishop, C. M. *Pattern recognition and machine learning*. Information science and statistics. Springer, New York, 2006.

[5] Bourdes, V., Bonnevay, S., Lisboa, P. J. G., Aung, M. S. H., Chabaud, S., Bachelot, T., Perol, D., and Negrier, S. Breast cancer predictions by neural networks analysis: a comparison with logistic regression. In *Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE* (Aug 2007), pp. 5424–5427.

[6] Bozek, J., Mustra, M., Delac, K., and Grgic, M. A survey of image processing algorithms in digital mammography. In *Recent Advances in Multimedia Signal Processing and Communications*, M. Grgic, K. Delac, and M. Ghanbari, Eds., vol. 231 of *Studies in Computational Intelligence*. Springer Berlin Heidelberg, 2009, pp. 631–657.

[7] Brandan, M. E. Detección del cáncer de mama:: Estado de la mamografía en méxico., 2006.

[8] Bull, L. *Applications of learning classifier systems*. Studies in fuzziness and soft computing,. Springer, Berlin ; New York, 2004.

[9] Campanini, R., and Lanconelli, N. *Support Vector Machines in CAD Mammography*. SPIE Press, 2006, ch. 7.

[10] Castro Astudillo, A. C., and Terashima Marín, H. *Detención, etiquetado y reconstrucción de masas y su clasificación por medio de redes neuronales en mamografías digitales.* 2012.

[11] Cheng, H., and Cui, M. Mass lesion detection with a fuzzy neural network. *Pattern Recognition 37*, 6 (2004), 1189.

[12] Cheng, H., Shi, X., Min, R., Hu, L., Cai, X., and Du, H. Approaches for automated detection and classification of masses in mammograms. *Pattern Recognition 39*, 4 (2006), 646 – 668.

[13] Cheng, L., and Li, X. Breast imaging reporting and datasystem (bi-rads) of magnetics resonance imaging: Breast mass. *Gland Surgery 1*, 1 (2012).

[14] Chhatwal, J., Alagoz, O., Lindstrom, M. J., Kahn, C. E., Shaffer, K. A., and Burnside, E. S. A logistic regression model based on the national mammography database format to aid breast cancer diagnosis. *American Journal of Roentgenology 192*, 4 (2009), 1117–1127.

[15] Conant-Pablos, S., Hernández-Cisneros, R., and Terashima-Marín, H. *Genetic and Evolutionary Computation: Medical Applications*. Wiley and Sons, ch. Feature Selection for the Classification of Microcalcifications in Digital Mammograms using Genetic Algorithms, Sequential Search and Class Separability.

[16] Conant-Pablos, S. E., Hernández-Cisneros, R. R., and Terashima-Marín, H. *Feature Selection for the Classification of Digital Mammograms using Genetic Algorithms, Sequential Search and Class Separability.* Genetic and Evolutionary Computation: Medical Applications. S. Smith and S. Cagnoni. Wiley, 2010.

[17] Consejo de Salubridad General. Diagnóstico y tratamiento del cÁncer de mama en segundo y tercer nivel de atención, 2009.

[18] Deserno, T. *Biomedical Image Processing*. Biological and Medical Physics, Biomedical Engineering. Springer, 2011.

[19] D'Orsi, C., and Newell, M. Bi-rads decoded: Detailed guidance on potentially confusing issues. *Radiologic Clinic of North America 45*, 5 (n.d.).

[20] Edén A. Alanís-Reyes, José L. Hernández-Cruz, H. T.-M. S. E. C.-P. Evolutionary feature selection applied to the classification of microcalcification clusters and masses for breast cancer computer-aided diagnosis. In *MIBISOC 2013, International Conference on Medical Imaging using Bio-inspired and Soft Computing* (2013), pp. 167–174.

[21] Franco Villalobos, C., and Tamez Peña, J. G. *Efecto de filtros digitales en la detención de biomarcadores en mamografías digitales / por Conrado Franco Villalobos ; [asesor, Dr. José Gerardo Tamez Peña].* 2012, 2012.

[22] Goldberg, D. E. *Genetic Algorithms in Search, Optimization and Machine Learning*, 1st ed. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1989.

[23] Gonzalez, R. C., and Woods, R. E. *Digital image processing / Rafael C. Gonzalez, Richard E. Woods.* Upper Saddle River, NJ : Pearson/Prentice Hall, c2008., 2008.

[24] Guyon, I., and Elisseeff, A. An introduction to variable and feature selection. *Journal of Machine Learning Research 3*, 7/8 (2003), 1157 – 1182.

[25] Hastie, T., Tibshirani, R., and Buja, A. Flexible discriminant analysis by optimal scoring. *Journal of the American Statistical Association 89* (1994), 1255–1270.

[26] Heath, M., Bowyer, K. W., and Kopans, D. Current Status of the Digital Database for Screening Mammography. Kluwer Academic Publishers, pp. 457–460.

[27] Heath, M. D., and Bowyer, K. W. Mass detection by relative image intensity. In *International Workshop on Database Machines* (2000).

[28] Hernandez-Cisneros, R., and Terashima-Marin, H. Evolutionary neural networks applied to the classification of microcalcification clusters in digital mammograms. In *Evolutionary Computation, 2006. CEC 2006. IEEE Congress on* (2006), pp. 2459–2466.

[29] Hernández Orallo, J., Ramírez Quintana, M. J., and Ferri Ramírez, C. *Introducción a la minería de datos / José Hernández Orallo, Ma. José Ramírez Quintana, Cèsar Ferri Ramírez.* Madrid : Pearson Educación : Prentice Hall, c2004., 2004.

[30] Huang, H., Wu, Y., Chan, Y., and Lin, C. Study on image feature selection: A genetic algorithm approach. In *Frontier Computing. Theory, Technologies and Applications, 2010 IET International Conference on* (Aug 2010), pp. 169–174.

[31] Hunt, K., Newman, L., Copeland, E., and Bland, K. *The Breast.*, 9â« ed ed. McGraw-Hill, New York, 2010, ch. 17.

[32] Instituto Nacional de Cancerología. Cáncer en cifras: Estadísticas 2008: Morbilidad, 2009.

[33] Knaul, F. M., Nigenda, G., Lozano, R., Arreola-Ornelas, Langer, A., and Frenk, J. Cáncer de mama en México: una prioridad apremiante. *Salud de México 51* (00 2009), s335 – s344.

[34] Kopans, D. B. *Breast imaging / Daniel B. Kopans.* Baltimore, MD : Lippincott Williams Wilkins, c2007., 2007.

[35] Kozlov, A., and Koller, D. Nonuniform dynamic discretization in hybrid networks. In *In Proceedings of the 13th Annual Conference on Uncertainty in AI (UAI)* (Providence, Rhode Island, 1997), Morgan Kaufmann, pp. 314–325.

[36] Kurkova, V. Kolmogorov's theorem and multilayer neural networks. *Neural Networks 5*, 3 (1992), 501–506.

[37] Llobet Azpitarte, R. Aportaciones al diagnóstico de cáncer asistido por ordenador. Master's thesis, 2006.

[38] M., A. K., and Sheshadri, H. S. On the classification of imbalanced datasets. *International Journal of Computer Applications 44*, 8 (April 2012), 1–7.

[39] McPhee, S. J., and Papadakis, M. A. *Current medical diagnosis and treatment 2009/ edited by Stephen J. McPhee, Maxine A. Papadakis.* New York : McGraw-Hill Companies, 2009. 48th ed., 2008.

[40] Mitra, S., and Acharya, T. *Data Mining. Multimedia, Soft Computing, and Bioinformatics.* Wiley, 2003.

[41] MM, P., and K., H. *The Breast.*, 10â« ed ed. McGraw-Hill, New York, 2007, ch. 63.

[42] Mohri, M., Talwalkar, A., and Rostamizadeh, A. *Foundations of Machine Learning.* Adaptive Computation and Machine Learning Series. MIT Press, 2012.

[43] National Cancer Institute. Mammograms: Fact sheet, 2010.

[44] O'Doherty, T. Review of the effective image processing techniques of mammograms.

[45] Oporto Díaz, S. A., and Terashima Marín, H. Detección automática de agrupamientos de microcalificaciones en mamografías digitalizadas / samuel alonso oporto díaz ; [asesor, hugo terashima marín]. Master's thesis, 2004.

[46] Organization, World Health. Globocan 2008: Cancer incidence, mortality and prevalence worldwide in 2008, 2009.

[47] Organization, World Health. Media centre. fact sheet no. 297: Cancer, 2009.

[48] Pitas, I. *Digital image processing algorithms and applications / I. Pitas.* New York : Wiley, c2000, 2000.

[49] Rojas-Domínguez, A., and Nandi, A. Development of tolerant features for characterization of masses in mammograms. *Computers in Biology and Medicine 39*, 8 (2009), 678–688.

[50] Russ, J. C. *The image processing handbook / John C. Russ.* Boca Raton : CRC/Taylor and Francis, c2007., 2007.

[51] Shen, Y., and Zelen, M. Screening sensitivity and sojourn time from breast cancer early detection clinical trials. mammograms and physical examinations. *Journal of Clinical Oncology 19*, 15 (2001), 3490 – 3499.

[52] Sickles, EA, D. C. B. L. e. a.

[53] Silverstein, M. J., and Recht, A. e. a. Image-detected breast cancer: State-of-the-art diagnosis and treatment. *Journal of the American College of Surgeons* (2009).

[54] Soille, P. *Morphological Image Analysis: Principles and Applications*. Springer, 2010.

[55] Steinwart, I., and Christmann, A. *Support vector machines*, 1st ed. Information science and statistics. Springer, New York, 2008.

[56] Suckling, J., Boggis, C., Hutt, I., Astley, S., Betal, D., Cerneaz, N., Dance, D., Kok, S., Parker, J., Ricketts, I., Savage, J., Stamatakis, E., and Taylor, P. The mammographic image analysis society digital mammogram database. *Exerpta Medica* (1994), 375–378.

[57] Tang, J., Rangayyan, R., Xu, J., El Naqa, I., and Yang, Y. Computer-aided detection and diagnosis of breast cancer with mammography: recent advances. 236–251.

[58] Turkington, C., and Krag, K. *Encyclopedia of Breast Cancer*. Facts on File Library of Health and Living, USA, 2005.

[59] Valenzuela-Rendón, M. Análisis de cruce multipunto y uniforme. Tech. rep.

[60] Valenzuela-Rendón, M. Clasicación de los métodos de selección. Tech. rep.

[61] Valenzuela-Rendón, M. Implementación de un algoritmo genético en una aplicaci ón. Tech. rep.

[62] Valenzuela-Rendón, M. The virtual gene genetic algorithm. In *Proceedings of the 2003 International Conference on Genetic and Evolutionary Computation: PartII* (Berlin, Heidelberg, 2003), GECCO'03, Springer-Verlag, pp. 1457–1468.

[63] Verma, B., McLeod, P., and Klevansky, A. Classification of benign and malignant patterns in digital mammograms for the diagnosis of breast cancer. *Expert Systems with Applications 37*, 4 (2010), 3344 – 3351.

[64] Zhang, Y., Tomuro, N., Furst, J., and Stan Raicu, D. Using bi-rads descriptors and ensemble learning for classifying masses in mammograms. In *Medical Content-Based Retrieval for Clinical Decision Support*, B. Caputo, H. MÃ$\frac{1}{4}$ller, T. Syeda-Mahmood, J. Duncan, F. Wang, and J. Kalpathy-Cramer, Eds., vol. 5853 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2010, pp. 69–76.

[65] Zhou, X., Liu, K.-Y., and Wong, S. T. Cancer classification and prediction using logistic regression with bayesian gene selection. *Journal of Biomedical Informatics 37*, 4 (2004), 249 – 259. Biomedical Machine Learning.

[66] Zyout, I., Abdel-Qader, I., and Jacobs, C. Embedded feature selection using pso-knn: Shape-based diagnosis of microcalcification clusters in mammography. *JUSPN* (2011), 7–11.

Vita

José Luis Hernández Cruz nació en la ciudad de Zacactecoluca, departamento de La Paz, El Salvador el 7 de mayo del año 1988. Obtuvo el título de Ingeniero en Tecnologías de Información y Comunicaciones por el Instituto Tecnológico y de Estudios Superiores de Monterrey, Campus Monterrey en diciembre de 2011. Fue admitido en el programa de posgrado de la Escuela de Ingeniería y Tecnologías de Información en diciembre de 2011 para la Maestría en Ciencias con especialidad en Sistemas Inteligentes.

José Luis Hernández Cruz was born in Zacatecoluca, El Salvador, on May 7, 1988. He earned the Information Technologies and Communications Engineering degree from the Instituto Tecnológico y de Estudios Superiores de Monterrey, Monterrey Campus in December 2011. He was accepted in the graduate programs in Information Technologies and Electronics in December 2011 for the Master of Science in Intelligent Systems.

La presente tesis fue tipografiada con LaTeX 2_ε por José Luis Hernández Cruz.