SEP
SECRETARÍA DE EDUCACIÓN PÚBLICA
CONACYT
Consejo Nacional de Ciencia y Tecnología
TECNOLÓGICO NACIONAL DE MÉXICO
Instituto Tecnológico de Ciudad Guzmán

**INSTITUTO TECNOLÓGICO DE CD. GUZMÁN**

PROGRAMA DE MAESTRÍA EN CIENCIAS DE LA COMPUTACIÓN

# TESIS

TEMA:

*"Multiclasificación de los niveles de densidad mamográfica con programación genética"*

**QUE PARA OBTENER EL GRADO DE:**

*MAESTRA EN CIENCIAS DE LA COMPUTACIÓN*

PRESENTA:

***Ing. Ninderlisbhet Vargas Cárdenas***

DIRECTORES:

*Dra. María Guadalupe Sánchez Cervantes*
*Dr. Daniel Fajardo Delgado*

CD.GUZMÁN, JALISCO, MÉXICO, AGOSTO, 2019

# Índice general

# Índice de figuras

# Índice de tablas

## Resumen

*El cáncer de mama es el más frecuente en las mujeres, al igual que el cervicouterino y el de piel. Las investigaciones médicas para la prevención del cáncer de mama han demostrado que la densidad mamaria es un fuerte indicador del riesgo de cáncer. La densidad puede evaluarse a través de la clasificación propuesta por el Colegio Americano de Radiología (ACR). La presente tesis tiene como objetivo mostrar los resultados obtenidos de la multiclasificación de los niveles de densidad mamográfica utilizando la programación genética. La clasificación multiclase parte de un conjunto de características de textura de las imágenes mamográficas. Se han implementado diversos métodos de extracción de características y de clasificación. Las características de las imágenes son la entrada para los clasificadores. Para los experimentos, se utilizó la base de datos de mamografías de INbreast. Los resultados muestran buena clasificación, donde se ha alcanzado un 70 % de efectividad de clasificación de los niveles de densidad mamográfica a través de la programación genética. Lo anterior se ejecuta en un ambiente paralelo que agiliza la obtención de resultados que servirán como apoyo a los médicos para realizar un diagnóstico fiable del riesgo de cáncer.*

## Abstract

*Breast cancer is the most common in women, as is cervical and skin cancer. Medical research for breast cancer prevention has shown that breast density is a strong indicator of cancer risk. Density can be assessed through the classification proposed by the American College of Radiology (ACR). This thesis aims to show the results obtained from the multiclassification of mammographic density levels using genetic programming. The multiclass classification is based on a set of texture features of mammographic images. Various methods of feature extraction and classification have been implemented. The features of the images are the input for the classifiers. For the experiments, the INbreast mammography database was used. The results show good classification, where 70 % effectiveness of classification of mammographic density levels has been achieved through genetic programming. The foregoing is carried out in a parallel environment that speeds up the obtaining of results that will serve as support for doctors to make a reliable diagnosis of cancer risk.*

# Autorización

EDUCACIÓN
SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLÓGICO NACIONAL DE MÉXICO

Instituto Tecnológico de Ciudad Guzmán

"2019, Año del Caudillo del Sur, Emiliano Zapata"

Cd. Guzmán, Jal. a 13/Agosto/2019

Oficio No. DEPI/50/19

ASUNTO : AUTORIZACIÓN DE IMPRESIÓN

**C. NINDERLISBHET VARGAS CÁRDENAS**
**N.C. M17290015**

En cumplimiento con el documento normativo de las disposiciones para la operación de estudios de posgrado del Tecnológico Nacional de México y con base en la aprobación del Comité Tutorial comisionado para su revisión; la División de Estudios de Posgrado e Investigación le otorga la autorización de impresión de su trabajo de tesis intitulado:

**"MULTICLASIFICACIÓN DE LOS NIVELES DE DENSIDAD MAMOGRÁFICA CON PROGRAMACIÓN GENÉTICA"**

dirigido por el **Dra. María Guadalupe Sánchez Cervantes**, desarrollado como requisito parcial para la obtención del grado de Maestro en Ciencias de la Computación, de acuerdo al plan de estudios MCCOM-2011-05.

Sin otro asunto en particular, quedo de usted.

**A T E N T A M E N T E**

**DR. HUMBERTO BRACAMONTES DEL TORO**
**JEFE DE LA DIVISIÓN DE ESTUDIOS DE POSGRADO E INVESTIGACIÓN**

S.E.P. TecNM
INSTITUTO TECNOLOGICO
DE CD. GUZMAN
DIVISION DE ESTUDIOS
DE POSGRADO E
INVESTIGACION

C.p. Archivo

# CAPÍTULO 1

# Introducción

## Contenido



El cáncer de mama es uno de los más comunes en mujeres, después del cáncer cervicouterino, y se prevee que para las próximas décadas aumente la mortalidad por este tipo de cáncer. Un análisis de los resultados de la detección del cáncer de mama expuesto por Knaul et. al.(2008), muestra como para la autodetección en la primera etapa de esta enfermedad solo el 10 % de los casos son detectables. Por otra parte, sólo el 22 % de las mujeres de 40 a 69 años informaron haber tenido detección de anomalías mamarias mediante una mamografía. El cáncer de mama puede ser detectado a través de diversas técnicas, mismas con las que se logra el diagnóstico de patologías como microcalcificaciones y densidades mamarias, sólo por mencionar algunas. De acuerdo a Arancibia et al. (2013) las calcificaciones mamarias corresponden a depósitos de calcio al interior del tejido mamario. Son hallazgos frecuentes en la mamografía, especialmente en mujeres postmenopáusicas. Si bien, la mayoría de las microcalcificaciones son originadas por patología benigna, algunos patrones agrupados específicos pueden ser causados por patología maligna. Aproximadamente el 55 % de los cánceres de mama no palpables presentan microcalcificaciones visibles.

Por otro lado, Saftlas et al. (1991) menciona que las densidades mamográficas son áreas del tejido mamario vistas radiográficamente por encima de la grasa. Las densidades están compuestas de tejido conjuntivo y epitelial, el tipo de tejido a partir del cual se desarrollan la mayoría de las neoplasias mamarias. Identificaron la relación de las densidades mamográficas con el riesgo posterior de cáncer de mama a través de un estudio de casos y controles utilizando

mamografías de prediagnóstico, logrando determinar que las probabilidades de cáncer de mama aumentaron constantemente con el aumento de la densidad mamaria. Es así, que se presenta la necesidad de clasificar las densidades mamarias y para ello existen varios métodos que facilitan el estudio de las mismas.

Por otra parte, se han implementado algunos métodos de clasificación de densidades mamarias, tales como el sistema "TNM" del AJCC (American Joint Committee on Cancer), y BIRADS (Breast Imaging Reporting and Data Systems). Sin embargo, la detección del cáncer de mama con los métodos convencionales aplicados por los especialistas no ha sido suficiente por lo que ahora se han implementado algunos métodos dinámicos.

En vías de brindar una solución al problema de clasificación, se ha hecho hincapié en diversas técnicas de clasificación, con base a ello Burling et al. (2016) utilizaron la Programación Genética (GP, por sus siglas en inglés), un paradigma de cómputo evolutivo para resolver el problema de clasificación binaria de niveles de densidad mamográfica. También hacen una comparativa con los Sistemas de Clasificación de Aprendizaje (LCS, por sus siglas en inglés). Es así que Burling y colaboradores, muestran como para un conjunto de datos MIAS e INBreas, la GP tiene mucho potencial para la clasificación de los niveles de densidad mamográfica.

Otros autores consideran diferentes estrategias para la resolución de la clasificación de los niveles de densidad mamográfica, es así que, autores como Smart (2005) hacen referencia a que la Programación Genética es implementada en resolución de tareas de clasificación de objetos de tres o más clases (multiclase). Así mismo, se han desarrollado métodos para mejorar el rendimiento de GP para clasificación de objetos multiclase, mediante la investigación de dos aspectos de GP. El primer aspecto de GP es la estrategia de clasificación, o el método utilizado para traducir un resultado real del programa en una etiqueta de clase para la clasificación, y en segundo, las estrategias de clasificación previas, mismas que suelen organizar el espacio de salida del programa en regiones de clase, que en algunos métodos pueden cambiar de posición o clase durante la evolución.

Ahora bien, es preciso reconocer el problema para abordarlo, mismo que se expresa a continuación.

En vías de revertir la tendencia creciente de mortalidad se ha incrementado el interés en el desarrollo de métodos computacionales para asistir en el análisis y diagnóstico de patologías a través de mamografías. Es preciso hacer mención de la existencia de métodos de clasificación

de niveles de densidad mamográfica en imágenes radiográficas digitales utilizando la Programación Genética y sistemas de clasificación de aprendizaje, mismos que se han expresado de manera limitada a otro tipo de problemas desarrollados con base a una clasificación por medio de árboles de decision. Ahora bien, parte de la propuesta es crear un multiclasificador a través de la Programación Genética. Para realizar la clasificación mediante GP, hay una serie de pasos a seguir, tales como: el pre-procesamiento de la imagen, la extracción de las características y la implementación de GP para la multiclasificación, donde el resultado de todo lo anterior en conjunto supone un considerable costo computacional. Por esta razón, se propone también la paralelización del método de multiclasificación de los niveles de densidad mamográfica.

## 1.1. Objetivos

A continuación se expresan los objetivos de la tesis.

### 1.1.1. Objetivo general

Generar un método de multiclasificación de los niveles de densidad mamográfica utilizando la Programación Genética.

### 1.1.2. Objetivos específicos

Para poder llevar a cabo el objetivo general, se plantean los siguientes objetivos específicos:

- Obtener una region de interés mediante el pre-procesamiento de la imagen mamográfica.
- Extraer las características a partir de la región de interés de la imagen mamográfica.
- Programar un método de multiclasificación de los niveles de densidad mamográfica con Programación Genética.
- Programar en un ambiente paralelo la multiclasificación de los niveles de densidad mamográfica y evaluar la mejora en términos de tiempo del programa en paralelo.

## 1.2. Preguntas de investigación

Se expresan preguntas de investigación para esta tesis, tales como las siguientes:

- ¿Resulta efectivo el pre-procesamiento de una imagen para la correcta extracción de características importantes?
- ¿El método de multiclasificación de los niveles de densidad mamográfica a través de la Programación Genética representa una mejora en exactitud respecto a otras técnicas de clasificación?
- ¿Se efectúa una reducción en términos de tiempo con el método paralelizado, en comparación con la aplicación secuencial?

## 1.3. Justificación

La presente investigación tiene como objetivo principal diseñar un método de multiclasificación de los niveles de densidad mamográfica utilizando la Programación Genética y debido al costo computacional que conlleva, también se implementa en un ambiente paraleo, lo que representará una reducción en tiempo computacional.

A continuación se menciona el impacto social, tecnológico, económico y ambiental que el presente trabajo de investigación tienen en cada uno de estos aspectos.

El impacto social se genera a partir del método de multiclasificación de los niveles de densidad mamográfica con Programación Genética, donde este funcionará de manera más rápida y efectiva en apoyo a la detección del cáncer de mama, debido a que el cáncer de mama es uno de los de mayor incidencia después del cáncer cervicouterino. Aunado a esto, se sabe que los casos detectados de cáncer de mama en una primera etapa oscila alrededor del 10 % y se estima que incremente sustancialmente en las próximas décadas. Es así que al mejorar el tiempo de ejecución de un método de clasificación de los niveles de densidad mamográfica con Programación Genética se prevee que tenga una mejor implementación y por ende contribuya a la efectiva multiclasificación de los niveles de densidad mamográfica para finalmente revertir la incidencia de cáncer de mama en los próximos años. El impacto

tecnológico se sustenta en que, existen pocos trabajos para la multiclasificación de densidades mamográficas con Programación Genética; de tal manera que la implementación de la paralelización de dicho método representa un aporte a los trabajos que antecede. El impacto económico se percibe desde la reducción del tiempo computacional con la paralelización del método de multiclasificación de los niveles de densidad mamográfica que generará un ahorro en tiempo computacional y por lo tanto un ahorro económico en energía. También, es importante mencionar que el impacto ambiental se dará desde el momento en que se realice la paralelización del método para la multiclasificación de los niveles de densidad mamográfica que podrá fomentar el ahorro de energía eléctrica lo que es equiparable a disminuir el consumo de los recursos para la generación de electricidad y por ende la emisión de gases contaminantes.

## 1.4. Organización de la tesis

La organización del presente documento queda estructurada de la siguiente manera:

El Capítulo 2 presenta el estado del arte donde se plasman los antecedentes del tema de investigación y también se enfoca en la fundamentación teórica de los tópicos inmersos en este trabajo, con la finalidad de mantener una comprensión de los conceptos que se manejarán a lo largo del extenso. El Capítulo 3 expone la extracción de características que parte del pre-procesamiento de la imagen, de tal manera que se haga una recopilación de características implementando diversos métodos de extracción en conjunto. También, se muestra el desarrollo del método de multiclasificación para los niveles de densidad mamográfica con Programación Genética. Así mismo, se aborda el programa genético de multiclasificación en un ambiente paralelo aplicando el modelo de paralelización por islas. El Capítulo 4 presenta los resultados obtenidos del trabajo de investigación respecto a los temas de multiclasificación y la paralelización. El Capítulo 5 muestran las conclusiones y trabajo futuro relacionado con la presente investigación. El capítulo 6 expone los productos obtenidos dentro del desarrollo de esta tesis. Para finalizar, en el Capítulo 7 se presenta un glosario de términos relacionados con éste trabaho de tesis.

# CAPÍTULO 2

# Fundamento teórico

## Contenido



Es importante mencionar que existen diversos trabajos que anteceden al presente, sin embargo, cada uno de ellos tiene una orientación específica como se podrá apreciar en este capítulo, así mismo la interpretación de la investigación se apoya en diversos conceptos como los que se mencionan a continuación, mismos que son importantes para la comprensión de esta tesis.

## 2.1. Antecedentes

Principalmente, es importante mostrar el trabajo activo respecto al tema de la multiclasificación para los niveles de densidad mamográfica y para ello, en la tabla 2.1 se exponen

algunos trabajos relacionados a la clasificación de los niveles de densidad mamográfica basados en el conjunto de datos MIAS, además, se presenta la cantidad de imágenes utilizadas, el número de clases a clasificar, tipo de ROI (Región de Interés, por sus siglas en inglés), los clasificadores implementados, el porcentaje de exactitud y algunas características específicas para cada trabajo.

| Trabajos | Imágenes | Clases | Tamaños del ROI | Clasificador | Exactitud (%) | Otros Parámetros |
|---|---|---|---|---|---|---|
| Blot L, Zwiggelaar (2001) | 265 | 3 | SBT | kNN | 63.0 | -Pre-procesamiento con GLCM<br>-K = 32<br>-Vector de características = 8 |
| Bosch A, Munoz X, Oliver A, Marti J (2006) | 322 | 3 | SBT | SVM | 91.3 | -Pre-procesamiento con GLCM y k-means<br>-SIFT con 35,000 descriptores y espacio entre parches M = 2<br>-Textons con parche N = 3<br>-Para bases de datos MIAS y DDSM |
| Muhimmah I, Zwiggelaar R (2006) | 321 | 3 | SBT | DAG-SVM | 77.5 | -Extracción por multiresolución<br>-Histograma de 1,280 |
| Subashini TS, Ramalingam V, Palanivel S (2010) | 43 | 3 | SBT | SVM | 95.4 | -Extracción por metodos estadísticos<br>-Se aplica normalización de características con valores de -1 a 1<br>-Implementación de sigmoidal en kernel<br>-Validación cruzada con 3-fold |
| Tzikopoulos SD, Mavroforakis ME, Georgiou HV, Dimitropoulos N, Theodoridis S (2011) | 322 | 3 | SBT | SVM | 84.1 | -Orientado a la selección del ROI<br>-Pre-procesamiento para identificación del pezón con NippleROI |
| Li JB (2012) | 42 | 3 | SBT | KSFD | 94.4 | -Extracción por métodos estadísticos<br>-Vector de 7 dimensiones |
| Mustra M, Grgic M, Delac K (2012) | 322 | 3 | 512 x 384 | IB1 | 82.0 | -Implementacion de validacion cruzada con 10-folds<br>-Extracción con GLCM |
| Silva WR, Menotti D (2012) | 320 | 3 | 300 x 300 | SVM | 77.1 | -Extracción por metodos estadisticos<br>-Parametron de regularización C = 213<br>-Se aplica normalización de características con valores de -1 a 1<br>-Implementacion de validacion cruzada con 10-folds |
| Kriti, Virmani J, Thakur S (2016) | 322 | 3 | 200 X 200 | SVM | 86.3 | -Extraccion por metodos estadisticos<br>-Implementacion de validacion cruzada con 10-folds |
| Kriti, Virmani J (2016) | 322 | 3 | 200 x 200 | SVM | 87.5 | -Pre-procesamiento con Laws' Mask<br>-Parametron de regularización C = 215<br>-Kernel = 24 |
| Sharma V, Singh S (2014) | 322 | 2 | 200 x 200 | SMO-SVM | 96.4 | -Implementación de AUC-ROC<br>-Alta sensitividad y especificidad se denota con valores cercanos a 1<br>-No discriminación con 0.5 |
| Sharma V, Singh S (2015) | 212 | 2 | 200 x 200 | kNN | 97.2 | -Extracción con GLCM, Laws' Mask y FPS para un vector 3-dimensional.<br>-K = 3 |
| Kriti, Virmani J, Dey N, Kumar V (2015) | 322 | 2 | 200 x 200 | SVM | 94.4 | -Extracción con Laws' Mask<br>-Parametron de regularización C = 215<br>-Kernel = 24 |
| Kriti, Virmani J (2015) | 322 | 2 | 200 x 200 | kNN | 95.6 | -Extracción con Laws' Mask 3, 5, 7 y 9<br>-K = 8 |
| Virmani J, Kriti (2015) | 322 | 2 | 200 x 200 | kNN | 96.2 | -Implementación de 2D-DWT para generación de 7 subimágenes<br>-Pruebas con k = {1,2,...,9,10} |
| Zemmal N, Azizi N, Dey N, Sellami M (2016) | 322 | 2 | 200 x 200 | SSVM | 94.4 | -Pruebas para vector de características de dimensión = 28 |

Tabla 2.1: Analogía sobre trabajos previos para la clasificación de densidades en mamografías

Todos los trabajos que se enlistan en la Tabla 2.1 están centrados en la extracción de las características más que en la clasificación, y se basan principalmente en el conjunto de datos MIAS y en algunos casos en DDSM, pero hasta el momento ninguno reporta resultados sobre el uso de la base de datos INbreast.

Investigaciones como la de Reyad et al. (2014), afirman que las mamografías se encuentran entre las imágenes médicas más difíciles de analizar. Esto se debe al bajo contraste y al nivel de intensidad de este tipo de imágenes, que pueden ser fácilmente afectadas. Por otra parte, experimentaron con varios métodos para la extracción de características entre ellos los métodos estadísticos, los patrones binario locales, la Transformada Wavelet Discreta (DWT, por sus siglas en inglés) y la Transformada de Contorno (CT, por sus siglas en inglés). Lo anterior con la finalidad de demostrar que los métodos propuestos para la extracción de características resultan efectivos para la clasificación del ROI, declarando que la implementación de los patrones binarios locales en conjunto con los métodos estadísticos representan al mejor resultado.

Por otro lado, Sheshadri y Kandaswamy (2007) implementaron un análisis de textura bajo seis características de acuerdo con el histograma de intensidad: la media, la desviación estándar, la suavidad, el tercer momento, la uniformidad y la entropía, esto para la extracción de características e implementación en la clasificación para niveles de densidad mamaria. Así mismo hacen referencia a la extracción de características bajo métodos estadísticos, estructurales y espectrales. En los métodos estadísticos la discriminación de texturas se basa en una matriz de coocurrencia del nivel de grises, mientras que los métodos estructurales implementan un elemento básico de textura para crear un patrón de textura más complejo basándose en reglas gramaticales. Por otra parte, los métodos espectrales transforman la imagen texturizada en dominio de la frecuencia.

En cuanto a la clasificación, Smart y Zhang (2003) investigaron y exploraron métodos dinámicos de clasificación, tales como, el método de Selección de Rango Dinámico Centrado (CDRS, por sus siglas en inglés) y el Método Selección de Rango Dinámico Ranurado (SDRS, por sus siglas en inglés). Así, se demostró que para un número de clases de objetos arregladas en orden aleatorio se aplicaban con mayor eficiencia los métodos dinámicos mencionados anteriormente, donde CDRS y SDRS, funcionaron mejor que el método SRS estático en la mayoría de las tareas de clasificación de objetos. Por lo que concluyen que, para problemas binarios o terciarios de clasificación de objetos con clases en orden natural se pueden aplicar con mejor eficiencia los métodos dinámicos CDRS y SDRS.

Además de los métodos de clasificación mencionados existen técnicas para la solución de problemas de clasificación, entre ellas está la Programación Genética, misma que se ha definido como una técnica de aprendizaje que ofrece un gran potencial para la creación de clasificadores, de acuerdo con Burling et al., (2016). También es reconocida por ser una técnica heurística muy flexible que permite la implementación de patrones complejos como los árboles de decisión, ejemplificándola en cualquier tipo de operación que pueda usarse dentro de la implementación y el conocimiento del dominio para el proceso de aprendizaje, tal como lo explican en su trabajo Espejo et al. (2010). Además, describen a la Programación Genética como una técnica evolutiva que puede ser adaptable en la evolución de los clasificadores para una construcción precisa y confiable. De igual forma, hacen referencia a la aplicación de la GP en la clasificación, esto de una forma estructurada destacando que, el procesamiento incluye características de selección y construcción, la extracción del modelo puede generarse a través de árboles de decisión, reglas de clasificación y funciones discriminantes, por mencionar algunas y finalmente un ensamble de clasificadores.

Remarcando el aporte de Burling y coautores, también se mencionan de forma detallada las utilidades de diferentes patrones binarios locales y características estadísticas usando la GP y LCS mediante cuatro técnicas de clasificación convencionales, tales como Bayes, árboles de decisión, vecino más cercano a K y máquinas de vector soporte. Es así que sostienen como para un conjunto de datos MIAS e INBreast los algoritmos evolutivos son idóneos para la clasificación de densidades mamográficas. Cabe mencionar que, los trab

A diferencia de la clasificación binaria a la cual hacen referencia los autores anteriores, existen otras técnicas de clasificación multiclase como la de Ingalalli et al. (2014), quienes presentan un esquema de GP basada en árboles y hacen referencia a que generalmente se utilizan árboles de análisis sintáctico para representar a los individuos, donde el nodo raíz y todos los demás nodos terminales pertenecen a un conjunto $O$ de operaciones predefinidas y los nodos hoja pertenecen al conjunto de atributos $A$ para un conjunto de datos dado. Al final del proceso de búsqueda, la solución está disponible en la raíz del mejor árbol, que es una función fácilmente interpretable que se utiliza para la tarea de clasificación. También, modificaron la representación de cada árbol de análisis agregando un nodo raíz $r$ de aridad $d(d \geq 1)$, de tal manera que, el nodo raíz $r$ tendrá $d$ ramas y cada una de ellas es un árbol de GP, así cuando termina la evolución los individuos aún se pueden evaluar en el nodo raíz $r$. La diferencia después de la evolución es que obtienen $d$ soluciones distintas para la clasificación, donde el valor de $d$ es independiente. Ahora que cada individuo se representa por $d$ funciones diferentes se mapea un espacio de solución $d - dimensional$ para cualquier individuo con un nodo raíz $r$ de aridad $d$.

Asimismo, el trabajo de Loveard y Ciesielski (2001) tuvo como objetivo explorar representaciones alternativas e implementaciones de problemas de clasificación de clase múltiple dentro del paradigma GP, y determinar cuál método de representación era más adecuado para realizar tareas de clasificación. Para ello, se analizaron cinco métodos, sin embargo, los resultados indicaron que la selección del rango dinámico es más apropiada para problemas binarios siendo capaz de producir clasificadores con mayor grado de precisión. Además, se menciona la posibilidad de utilizar el método de selección de rango dinámico en conjuntos de datos de clase múltiple una vez que estos estén descompuestos en un subconjunto de datos binarios. Lo anterior con la finalidad de mejorar la aplicabilidad del método de selección de rango dinámico en problemas de clase múltiple.

En contraste, Smart (2005) investiga el uso de la Programación Genética en la resolución de tareas de clasificación de objetos de tres o más clases (multiclase). Se han desarrollado métodos para mejorar el rendimiento del sistema GP en cuatro tareas de clasificación de objetos multiclase, mediante la investigación de dos aspectos de GP. El primer aspecto de GP es la estrategia de clasificación, o el método utilizado para traducir un resultado real del programa en una etiqueta de clase para la clasificación. Las estrategias de clasificación previas suelen organizar el espacio de salida del programa en regiones de clase, que en algunos métodos pueden cambiar de posición o clase durante la evolución.

Por otro lado, algunos trabajos centrados en el cómputo paralelo enfocado a la ágil ejecución de tareas, se resalta el de Romero-Laorden et al. (2016) que habla acerca de la técnica beamforming (conformación de haces) y un algoritmo de selección de características del Método de Enfoque Total (TFM, por sus siglas en inglés) que usan señales adquiridas independientemente de pares de elementos de matriz de emisión-recepción. Este proceso produce imágenes de alta calidad con todos sus puntos enfocados tanto en la emisión como en la recepción, pero el costo computacional es alto. Por eso asegura que, es importante reconocer que los procedimientos que conllevan a la detección sistematizada del cáncer de mama (en este caso para la clasificación de los niveles de densidad mamográfica) pueden ser mejorados al involucrar técnicas de paralelización.

El trabajo de Melab et al. (2010) habla del modelo de isla paralela en GPU, así como su rediseño, implementación y problemas asociados relacionados con el contexto de ejecución de la GPU. Los resultados preliminares que repotan, demuestran la efectividad de los enfoques propuestos y sus capacidades para explotar completamente la arquitectura de la CPU y la GPU. También se centran en el modelo cooperativo sincrónico de la isla, y se despliegan simultáneamente para cooperar en el cálculo de mejores y más robustas soluciones. Intercam-

bian de forma genérica una manera sincrónica para diversificar la búsqueda. El objetivo es permitir el retraso de la convergencia global, especialmente cuando los Algoritmos Evolutivos (AE) son heterogéneos con respecto a los operadores de variación.

Chow et al. (2006) consideran principalmente la paralelización de los algoritmos estándar multigrid y mencionan el desarrollo de métodos eficaces en esta tecnología que se reduce a lograr un buen equilibrio entre los tiempos de configuración, las tasas de convergencia y el costo por iteración. Estas características a su vez dependen de la complejidad del operador, las tasas de crecimiento y una efectividad más suave. Es así que profundizan en simulaciones basadas en Ecuaciones Diferenciales Parciales (PDE, por sus siglas en inglés) mismas que están paralelizadas dividiendo el dominio de interés en subdominios (uno para cada procesador). Cada procesador es responsable de actualizar las incógnitas asociadas dentro de su subdominio solamente. Además, mencionan que el objetivo general es asignar a cada procesador una cantidad igual de trabajo y para minimizar la cantidad de comunicación entre procesadores minimizan esencialmente el área de superficie de los subdominios.

## 2.2. Cáncer de mama

El cáncer de mama es una enfermedad con una evolución natural compleja por lo que, a pesar de los avances de la oncología moderna, es la segunda causa de muerte por neoplasia en la mujer en el ámbito mundial, con cerca de 500 mil muertes cada año, de las cuales el 70 % ocurre en países en desarrollo según Cárdenas et al. (2013), quien también menciona que la detección actualmente se genera con el diagnóstico por imagen, que permite visualizar anomalías que exponen los riesgos detectables más comunes que son las microcalcificaciones y las densidades mamográficas, que podrán detectarse por diversas técnicas como la mastografía, el ultrasonido mamario, una resonancia magnética o por tomografía por emision de positrones. Las técnicas de diagnóstico por imagen se comprende por:

1. Mastografía. Es el único método de imagen que ha demostrado disminución en la mortalidad por cáncer de mama al permitir un diagnóstico temprano.
2. Ultrasonido (US) mamario. Valiosa herramienta complementaria de la mastografía diagnóstica, no útil como método de tamizaje para cáncer. Se requieren equipos de alta resolución, así como experiencia y conocimiento de la anatomía de la glándula mamaria y su evaluación por ecografía.

3. Resonancia magnética (RM). Otro método de imagen que no utiliza radiación ionizante y proporciona información morfológica y funcional, a través de la inyección endovenosa de una sustancia paramagnética (gadolinio).

4. Tomografía por emisión de positrones (PET CT). Es un estudio que combina tomografía computada (CT) con medicina nuclear (PET) en una misma imagen y permite en forma simultánea un estudio no sólo morfológico sino también funcional (metabólico) para la localización exacta de metástasis.

### 2.2.1. Niveles de densidad mamográfica

La densidad mamaria, estudiada a través de la mamografía (densidad mamográfica), refleja la composición del tejido mamario. El epitelio y estroma mamario producen mayor atenuación de los rayos X que la grasa, por lo que aparecen blancos en la mamografía, mientras que la grasa se ve oscura. Así, la apariencia de la mamografía varía entre las mujeres, dependiendo de la composición de su mama. La proporción de mama constituida por tejido conectivo y epitelial es usualmente denominada como porcentaje de tejido mamario o Porcentaje de Densidad Mamográfica (PDM), así como lo describe Neira (2012).

### 2.2.2. Tipos de clasificación de cáncer de mama

El método de clasificación más común para clasificar el cáncer de mama, es el sistema "TNM" del AJCC (American Joint Committee on Cancer).

- T se refiere al tamaño del tumor. El valor de T es de entre 0 y 4, y describe el tamaño del tumor y la extensión a la piel o la pared torácica. Los valores de T más elevados indican la presencia de un tumor mayor o una mayor extensión a los tejidos cercanos a la mama.

- N se refiere a los ganglios linfáticos. Los valores para N se sitúan entre 0 y 3, e indican si el cáncer se ha extendido a los ganglios linfáticos cercanos a la mama y, en caso de haberlo hecho, cuántos se encuentran afectados.

- M se refiere a la metástasis. Los valores de M son 0 (ausencia de metástasis) ó 1 (presencia de metástasis), e indican si el cáncer se ha extendido a otros órganos distantes como los pulmones o los huesos.

Existe un método de clasificación denominado BIRADS (Breast Imaging Reporting and Data Systems) desarrollada por el Colegio Americano de Radiología (1993), esta es implementada para la detección de microcalcificaciones y densidades mamarias, un ejemplo sería como el que se muestra en la Tabla 2.2 que explica la categoría con base a su especificación o incidencia y las recomendaciones médicas propuestas para cada tipo de hallazgo.

| **Categoría** | **Especificación** | **Recomendaciones** |
|---|---|---|
| 0 | -Insuficiente para diagnóstico<br>-Existe 13 % de posibilidad de malignidad | Se requiere evaluación con imágenes mamográficas. |
| 1 | -Negativo<br>-Ningún hallazgo que reportar | Mastografía de rutina anual para mujeres a partir de los 40 años de edad. |
| 2 | -Hallazgos benignos | Mastografía de rutina anual para mujeres a partir de los 40 años de edad. |
| 3 | -Hallazgos probablemente benignos<br>-Menos del 2 % de probabilidad de malignidad. | Requiere seguimiento por imagen del área con hallazgos de manera periódica. |
| 4 | -Hallazgos de sospecha de malignidad<br>-Se subdivide en:<br>4a - Baja sospecha de malignidad<br>4b - Sospecha intermedia de malignidad<br>4c - Hallazgos moderados de sospecha de malignidad pero no clásicos. | Requiere biopsia. |

Tabla 2.2: Clasificación BIRADS por el Colegio Americano de Radiología (1993)

## 2.3. Clasificación binaria y multiclasificación

De acuerdo con Frank y Hall (2001), la clasificación binaria está basada en algoritmos de clasificación que asignan un conjunto de etiquetas a un valor objetivo categórico bifurcado y hacen mención de la constante implemenación del aprendizaje automático, debido a que este a menudo involucra situaciones de orden entre varias categorías. Sin embargo Aly (2015) menciona que cuando se tienen clases múltiples, el problema principal puede descomponerse en varias tareas de clasifiación binaria y es una de las formas comunes de resolver problemas de clasificación multiclase. Aunado a esto, existen dierentes métodos de clasificación multiclase

como Uno Versus Todos (OVA, por sus siglas en inglés), Todos Versus Todo (AVA, por sus siglas en inglés), Corrección de Errores de Codificación de Salida (ECOC, por sus siglas en inglés) y la Codificación Generalizada.

## 2.4. Clasificadores

Actualmente, existen diversos clasificadores mismos que cumplen la función de asignar a un objeto clasificado una etiqueta de clase por diferentes procesos. A continuación se muestra una breve descripción de algunos clasificadores que, con base a los antecedentes se tomaron como referencia comparativa dentro de esta tesis, entre ellos están: Nearest Neighbor, Decision Tree, Random Forest, Support Vector Machine (SVM), Naive Bayes y Programación Genética.

### 2.4.1. Nearest neighbor

Nearest Neighbor, de acuerdo con Keller et al. (1985), es utilizado en su mayoría para el reconocimiento de patrones y consiste en asignar una característica como una función de distancia de un vector compuesto por los vecinos más cercanos a $k$, respecto a la característica de esos vecinos respecto a una clase. Así mismo, presenta una dificultad al momento de dar importancia a cada una de las muestras, esto debido a que, les asigna el mismo valor de relevancia.

### 2.4.2. Decision tree

Tal como lo menciona Swain y Hauska (1977), un árbol de decisión está caracterizado por la clasificación de muestras desconocidas a través de funciones de decisión. También describen que, un árbol de decisión está compuesto por un nodo raíz, varios nodos interiores y varios nodos terminales, de ellos el nodo raíz y los nodos interiores tienen la capacidad de vincularse con las etapas de decisión, mientras que los nodos terminales representan clasificaciones y valores finales.

### 2.4.3. Random forest

Bajo la definición de Breiman (2001), Random Forest representa una combinación de predictores de árboles, de tal manera que, cada árbol depende de los valores de un vector aleatorio muestreado de forma independiente y con la misma distribución para todos los árboles en el bosque.

### 2.4.4. Support vector machine

Cristianini y Shawe-Taylor (2000) mencionan que las SVM son un método de aprendizaje automático para la clasificación, la regresión y otras tareas de aprendizaje. Así mismo, mencionan que su funcionalidad consiste en mapear los vectores de entrada en el espacio de características de alta dimensión a través de un mapeo no lineal que es elegido a priori.

### 2.4.5. Naive bayes

Naive Bayes, de acuerdo con Zhang (2004), es uno de los algoritmos de aprendizaje inductivo más eficientes y efectivos para el aprendizaje automático y la minería de datos. La representación se basa en que, todos los atributos son independientes dado el valor de una variable de clase. También menciona que, para superar la limitación de los Naive Bayes es posible extender su estructura para representar explícitamente las dependencias entre los atributos.

## 2.5. Programación genética

La GP expresada por Páez (2001), forma parte de los Algoritmos Genéticos (AG), en la cual los cromosomas que sufren la adaptación son en sí mismos programas de computadora. Se usan operadores genéticos especializados que generalizan la recombinación y la mutación para los programas de computadora estructurados en árbol que están bajo adaptación. La GP trata de resolver uno de las cuestiones más excitantes e interesantes de las ciencias de la computación: ¿cómo pueden aprender las computadoras a solucionar problemas sin que se les

programe explícitamente? En otras palabras, la cuestión es cómo se puede hacer para que los computadores hagan lo que tienen que hacer, sin necesidad de la intervención humana que les diga exactamente como lo deben hacer.

La GP está basada e inspirada en la teoría darwiniana del evolución, por lo cual se constituye a partir de los siguientes aspectos:

- La población que representa al conjunto de individuos que a su vez son soluciones candidatas para la resolución de problemas.
- El método de herencia por generación que se implementa con la evolución.
- El método de selección (Fitness) de las mejores soluciones candidatas basadas en el espacio de búsqueda.
- La probabilidad de cruzamiento entre los individuos, que consiste básicamente en la selección de un punto aleatorio en cada uno de los dos individuos a cruzar y unicaménte intercambiar esa característica entre ellos.
- La probabilidad de mutación de un individuo, donde de un sólo individuo se selecciona un punto aleatorio que será modificado sin seguir algún patrón definido, pero respetando la estructura del árbol.

## 2.6. Cómputo paralelo

Turing (1950) fue el visionario que propuso el modelo de cómputo secuencial mediante el cual una máquina compuesta de una cabeza lectora, una cinta infinitamente larga y un conjunto finito de estados, podía resolver un problema lógico, de manera algorítmica a lo que se define como el nacimiento del computador secuencial, mismo que llamó Máquina de Estado Finito (MEF) que de forma evidente implementa el cómputo secuencial por la forma de resolver el problema lógico. Con la creciente necesidad del procesamiento ágil de grandes volúmenes de información, Sierra (2011) define el cómputo paralelo como una forma de procesamiento que permite dividir una tarea en varias partes, de esta forma cada una de ellas se procesa de forma simultánea, utilizando hardware con capacidades suficientes más no potentes, mejorando el tiempo de respuesta del procesamiento.

Dentro de los sistemas de cómputo con varios CPU, se ha tomado como base la taxonomía de Flynn (1972), quien eligió dos características consideradas esenciales: el número de flujos de instrucciones y el número de flujos de datos. "Una computadora con un flujo de instrucciones y uno de datos se llama SISD (Single Instruction Single Data), todas las computadoras tradicionales de un procesador (es decir aquellas que tienen un CPU) caen dentro de esta categoría desde las computadoras personales hasta las grandes mainframes. La siguiente categoría es SIMD (Single Instruction Multiple Data) con un flujo de instrucciones y varios flujos de datos. Este tipo se refiere a ordenar procesadores con unidad de instrucción que busca una instrucción y después instruye a varias unidades de datos para que la lleven a cabo en paralelo cada una con sus propios datos. Estas máquinas son útiles para los cómputos que repiten los mismos cálculos en varios conjuntos de datos. Por otra parte, está la categoría MISD (Multiple Instruction Single Data) con un flujo de varias instrucciones y un flujo de datos. Ninguna de las computadoras conocidas se ajusta a este modelo y por último está MIMD (Multiple Instruction Multiple Data) que significa un grupo de computadoras independientes cada una con su propio contador del programa y datos. Todos los sistemas distribuidos son MIMD" (Tanenbaum, 1996).

Un concepto importante sobre paralelismo es la Programación Dinámica (DP , por sus siglas en inglés) que se define según Kumar et al. (1994) como una técnica comúnmente utilizada para resolver una amplia variedad de problemas discretos de optimización, como programación, edición de cadenas, empaquetado y administración de inventario. DP ve un problema como un conjunto de subproblemas interdependientes. Resuelve subproblemas y usa los resultados para resolver subproblemas más grandes hasta que se resuelve todo el problema. A diferencia de Dividir-y-vencer, donde la solución a un problema depende solo de la solución a sus subproblemas, en DP puede haber interrelaciones entre los subproblemas. En DP, la solución a un subproblema se expresa como una función de soluciones a uno o más subproblemas en los niveles precedentes.

Otro término importante es el costo computacional que de acuerdo con Cormen et al, (2009), menciona que existen problemas capaces de resolverse en tiempo polinomial, es decir, para entradas de tamaño $n$, su peor tiempo de ejecución es $O(n^k)$ para una constante $k$. Pero, no todos los problemas se pueden resolver en tiempo polinomial, sin embargo, es una clase interesante de problemas, llamados problemas NP-completos cuyo estado es desconocido. Aún no se ha descubierto ningún algoritmo de tiempo polinomial para un problema NP-completo, ni nadie ha podido probar que no pueda existir un algoritmo de tiempo polinomial para ninguno de ellos.

Por otra parte, Beekman (2005) menciona que la velocidad de procesamiento está ligada con la función del procesador, mismo que está formado por un conjunto de registros que almacenen datos, una unidad aritmético-lógica que realiza operaciones con ellos y una unidad de control que se encarga de coordinar a todos los componentes. Debido a que las operaciones dentro del procesador se sincronizan con los tics del reloj, la velocidad máxima del ordenador vendrá marcada por el ritmo de oscilación del reloj interno.

Algunas tecnologías y arquitecturas de cómputo paralelo son las siguientes:

- SCOOP (Scalable Concurrent Operations in Python): Es un módulo que pertenece al lenguaje de programación Python especializado en la programación paralela concurrente en varios entornos con tareas distribuidas. "Scoop está basado en la filosofía de que el futuro es paralelo, lo simple es hermoso y el paralelismo debería ser más simple"(Hold-Geoffroy et al. 2014).
- CUDA (Compute Unified Device Architecture): Es una arquitectura de cálculo paralelo de NVIDIA que proporciona un incremento extraordinario del rendimiento de los sistemas a través del uso específico de la GPU. "Los sistemas informáticos están pasando de realizar el procesamiento central en la CPU a realizar coprocesamiento repartido entre la CPU y la GPU"(NVIDIA, 2017). Así mismo, NVIDIA ha inventado la arquitectura de cálculo paralelo CUDA, la cual incluye en sus GPUs GeForce, ION Quadro y Tesla GPUs, esto con la finalidad de hacer posible un procesamiento ágil, para el mejor rendimiento de los sistemas.
- MPI (Message Passing Interface): Es una interfaz estandarizada para la realización de aplicaciones paralelas basadas en paso de mensajes, misma interfaz que ha sido diseñada para sacar el mejor provecho a las arquiecturas con múltiples procesadores, tal como dice Alonso (1997), quien a su vez, habla sobre el modelo de programación que surge tras MPI, el cual es MIMD aunque se dan especiales facilidades para la utilización del modelo SPMD, un caso particular de MIMD en el que todos los procesos ejecutan el mismo programa, aunque no necesariamente la misma instrucción al mismo tiempo.
- OpenMP (Open Multi-Processing): Es una API que está centrada en la paralelización dirigida por el usuario, es decir, el programador dicta de forma explícita las acciones a realizar por el compilador y las directivas de tiempo de ejecución para ejecutar el programa en paralelo. De acuerdo con OpenMP (2015), las implementaciones compatibles con OpenMP no son necesarias para verificar dependencias de datos, conflictos de

datos, condiciones de carrera o interbloqueos, condiciones que facilitan la dirección del usuario sobre la paralelización.

## 2.7. Modelo de isla

El modelo de la isla según Latter (1973), parte de una especie que tiene la capacidad de subdividirse en varias poblaciones finitas entre las cuales se espera una migración de acuerdo con la magnitud de la misma. Sin embargo, no impone restricción alguna en la magnitud de las tasas de cruzamiento o mutación involucradas. También presenta como las propiedades de dos medidas fundamentales de divergencia genética se deducen de la teoría aplicativa sobre los parámetros $\mu$, el coeficiente de parentesco, y otro, $\lambda$, mide la tasa de divergencia mutacional entre las subpoblaciones.

"En una implementación paralela de un modelo de isla, cada máquina ejecuta un algoritmo genético y mantiene su propia subpoblación para la búsqueda"(Whitley, 1998). Es decir, las máquinas trabajan de manera cooperativa intercambiando regularmente parte de su población generando una migración donde se considera el intervalo de migración y las generaciones entre cada migración, así como la tasa de individuos a migrar. De igual forma, Whitley (1998) hace referencia a que los modelos de islas paralelas muestran un mejor rendimiento de búsqueda que los modelos de población única en serie, esto desde la calidad de la solución encontrada hasta el esfuerzo medido en la cantidad de evaluaciones en el espacio de búsqueda.

# CAPÍTULO 3

# Marco metodológico

## Contenido



En este capítulo se aborda el marco metodológico, así como la explicación de cada una de las etapas que contiene la metodología implementada. En la Figura 3.1 se muestra el esquema mencionado.

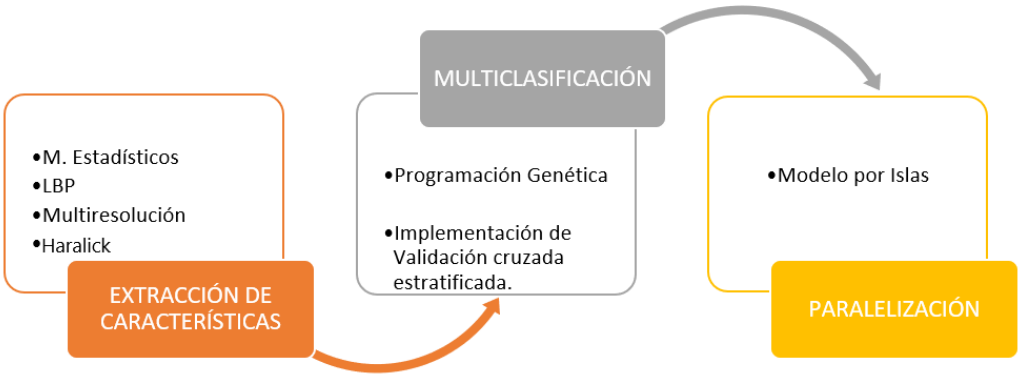


Figura 3.1: Esquema de la metodología implementada

## 3.1. Extracción de características

El primer paso para la multiclasificación efectiva de los niveles de densidad mamográfica es la correcta obtención de características que sigue una secuencia tal como: el pre-procesamiento de la imagen, la implementación mixta de un conjunto de métodos de extracción y la creación de un archivo para el almacenamiento de las características. De manera gráfica, en la Figura 3.2 se muestra la metodología de extracción de características que será explicada en la sección 3.1.2.

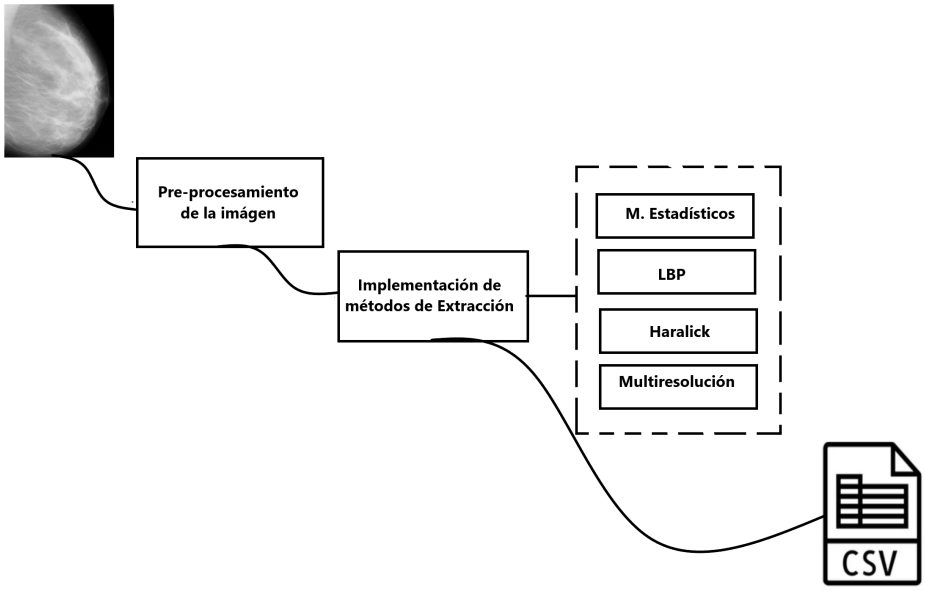


Figura 3.2: Metodología de extracción de características

### 3.1.1. Pre-procesamiento de las imágenes mamográficas

El pre-procesamiento de la imagen tiene la finalidad de obtener una región de interés (ROI, por sus siglas en inglés) que acote el campo de búsqueda para la eficiente extracción de las características y sigue un proceso jerárquico como el de la Figura 3.3

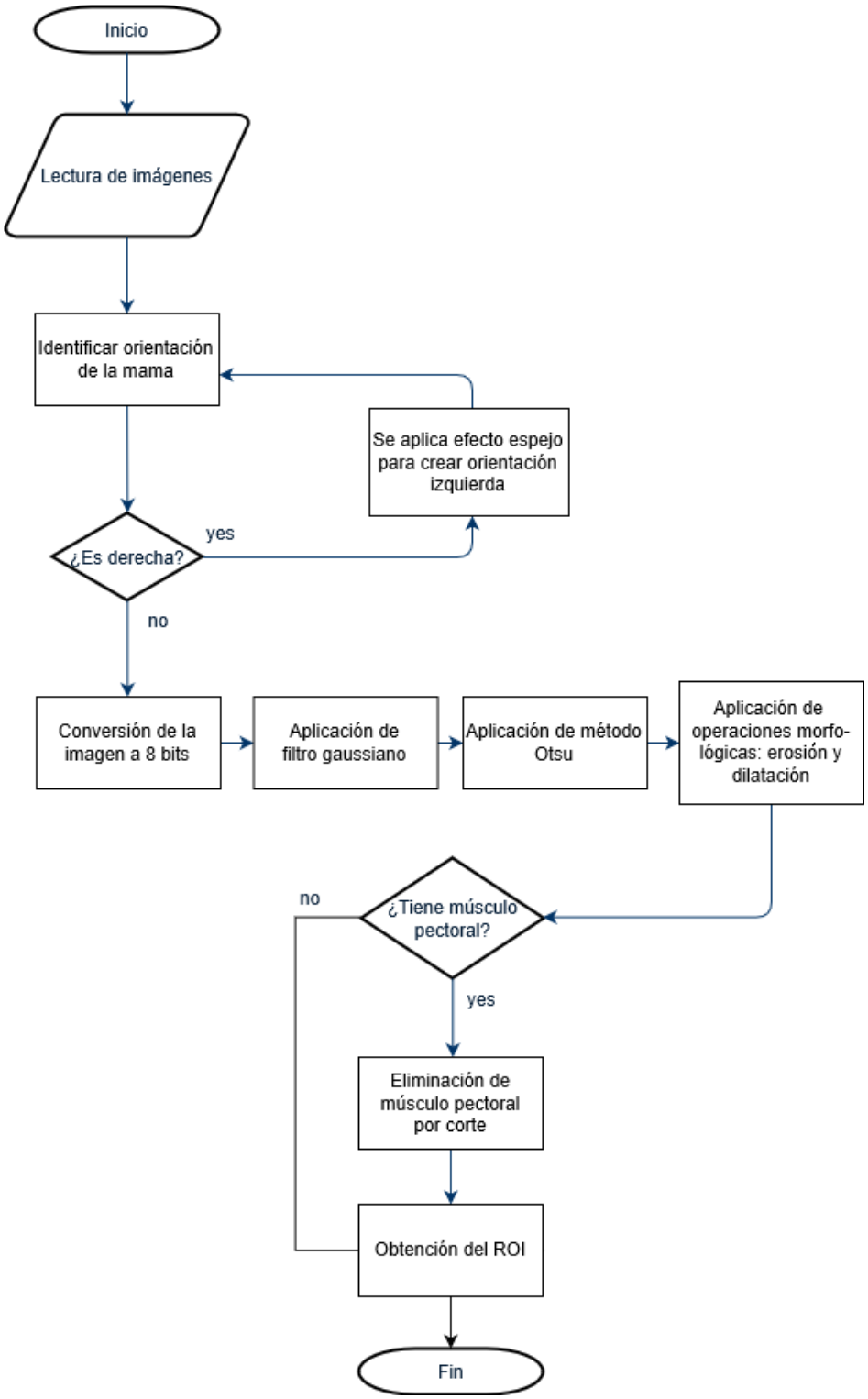


Figura 3.3: Diagrama de flujo del pre-procesamiento de imágenes mamográficas

El pre-procesamiento inicia con la lectura de cada una de las imágenes mamográficas, después se realiza la identificación de aquellas que tienen una alineación derecha de la mama, a las cuales se le aplica un efecto espejo que le da a la mama una alineación izquierda dentro de la imagen mamográfica, esto para el manejo estandarizado de la imagen, así como se muestra en la Figura 3.4.

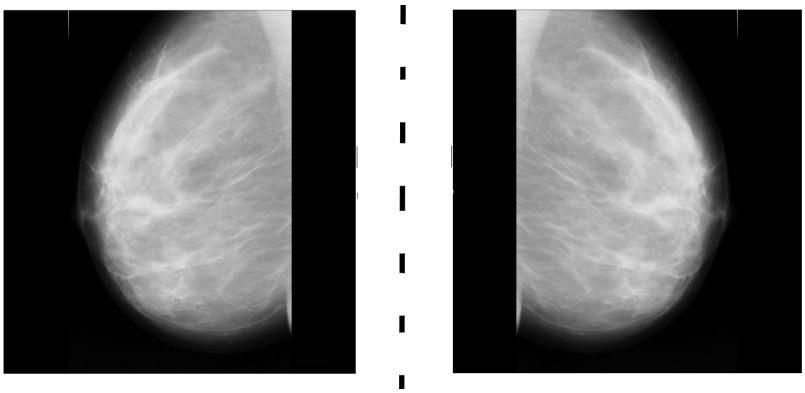

Figura 3.4: Efecto espejo de las imágenes mamográficas

Posteriormente, la imagen de 12 bits se convierte a una imagen de 8 bits, condición que permite manipular la imagen de tal manera que sea posible aplicar diversos filtros; uno de ellos es el filtro Gaussiano que muestra un aspecto de desenfoque o mejor conocido como suavidad, esto a través de una función gaussiana que reduce los píxeles de alta frecuencia de la imagen. La Figura 3.5 muestra un ejemplo cuando se aplica el filtro Gaussiano.

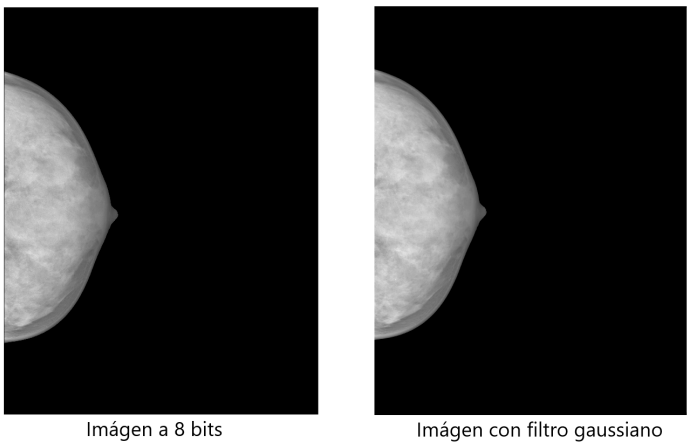


Figura 3.5: Imagen a 8 bits e imagen con filtro gaussiano

En seguida, se crea una máscara binaria, misma que se obtiene aplicando el método de Otsu (1979). Esta máscara se define a partir de un umbral, con el único objetivo de recuperar el área de la mama y descartar el fondo negro de las imágenes mamográficas, tal como se muestra en la Figura 3.6.

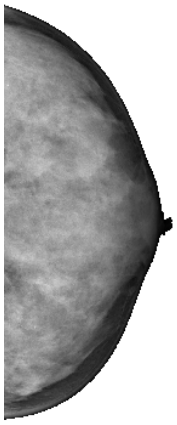

Figura 3.6: Máscara de la imagen

Una vez que se tiene la máscara binaria, se le aplican las operaciones morfológicas de erosion y dilatación. Estas funciones en conjunto tienen la finalidad de eliminar puntos externos a la región de interés para evitar que dichos generen ruido en la imagen. Dentro del análisis manual de las imágenes mamográficas, se logra identificar que en su mayoría estas muestran, además de la mama una sección del músculo pectoral, por lo que se identifican aquellas imágenes con esta característica y se aplica un corte arbitrario conveniente partiendo del centro de la mama para lograr descartar el músculo pectoral y obtener un ROI como se ejemplifica en la Figura 3.7.

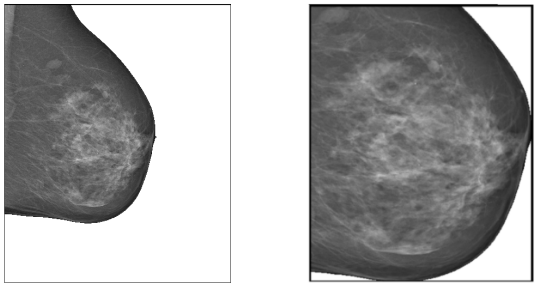


Figura 3.7: Imagen con músculo pectoral e imagen sin músculo pectoral

### 3.1.2. Métodos de extracción de características

La extracción de características forma parte fundamental del proceso de multiclasificación, debido a que impacta directamente con los resultados esperados. Es por ello que se toma en cuenta la composición de una imagen con base a sus características y, de forma específica, su textura por región que tiene la finalidad de especificar la parte de donde se van a extraer las características. La extracción de características está basada en las utilidades de implementar diferentes patrones locales binarios (LBP, por sus siglas en inglés) en conjunto con métodos estadísticos, el análisis multiresolución y el método de Haralick para los conjuntos de datos MIAS e INBreast.

#### Patrones Binarios Locales (LBP)

De acuerdo con Ojala et al. (2002), LBP es una característica de textura local que se ha aplicado en el reconocimiento de patrones debido a su alta potencia discriminativa. El histograma del ROI se calcula para cada píxel haciendo una comparativa entre los valores de nivel de gris con los de sus vecinos, y finalmente los histogramas se concatenan en forma de vector de características. Siguiendo este enfoque, se define $c$ como un píxel de la imagen mamográfica y $P$ un conjunto de píxeles vecinos a $c$ con un radio de distancia $R$. Teniendo en cuenta lo anterior se calcula el valor LBP de $c$ como se muestra en la siguiente ecuación (3.1) que compara el nivel de gris de $c$ expresado como $gc$, con el nivel de gris del conjunto

de píxeles vecinos $gP$. También en la ecuación (3.1), $s(x) = 1\ si\ x \geq 0,\ y\ s(x) = 0\ si\ x < 0$.

$$LBP_{P,R} = \sum_{P=0}^{P-1} s(gP - gc)2^P \tag{3.1}$$

.

Suponiendo que $c$ tiene las coordenadas (0,0), entonces las coordenadas de $P$ son $((-R\sin(\frac{2\pi P}{P}))$, $(R\cos(\frac{2\pi P}{P})))$. La Figura 3.8 muestra un ejemplo del conjunto de píxeles vecinos para diferentes valores de P y R.

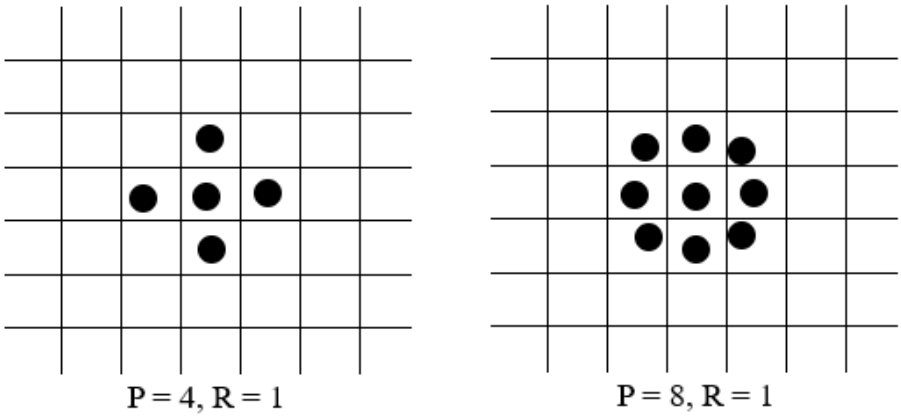


Figura 3.8: Conjunto de píxeles vecinos para diferentes valores de P y R

Para esta investigación el cálculo de los LBP se hace para cada píxel $c_{i,j}$ con coordenadas $i$, $j$ dentro de una imagen mamográfica de tamaño $N\ X\ M$. Después, se construye un histograma basado en la Ecuación 3.2 donde $K$ es el valor máximo del código LBP.

$$H(k) = \sum_{i=1}^{N}\sum_{j=1}^{M}(LBP_{P,R}\,(c_{i,j}\,), k \in [0, K]) \tag{3.2}$$

### Métodos estadísticos

Los métodos estadísticos son también implementados para describir la textura en el reconocimiento de patrones, y de acuerdo con Sheshadri et al. (2007) se basan en seis características estadísticas (media, desviación estándar, suavidad, asimetría, uniformidad y entropía). La

media es una medida de intensidad promedio, la desviación estándar es una medida de contraste promedio, la suavidad mide la suavidad relativa de la intensidad en una región, la asimetría mide la asimetría del histograma, la uniformidad mide la uniformidad de intensidad en el histograma y la entropía es una medida de aleatoriedad, tal como se describe en la Tabla 3.1. La expresión de los momentos de $n$-ésimo orden sobre la media es $\mu_n = \sum_{i=0}^{L-1}(Z_i - m)^m p(z_i)$, donde $Z_i$ es una variable aleatoria que indica la intensidad, $p(z_i)$ es el histograma de los niveles de intensidad en una región, $L$ es el número de los posibles niveles de instensidad, y $m$ es la intensidad media.

| Momento | Expresión | Medida de textura |
|---|---|---|
| Media | $m = \sum_{i=0}^{L-1} z_i p(z_i)$ | Medida de intensidad promedio |
| Desviación estárdar | $\sigma = \sqrt{\mu_2(z)} = \sqrt{\sigma^2}$ | Medida de contraste promedio |
| Suavidad | $R = 1 - \frac{1}{1+\sigma^2}$ | Suavidad de intensidad en una región |
| Asimetría | $\mu_3 = \sum_{i=0}^{L-1}(Z_i - m)^3 p(z_i)$ | Asimetría del histograma |
| Uniformidad | $U = \sum_{i=0}^{L-1} p^2(z_i)$ | Uniformidad de intensidad en histograma |
| Entropía | $e = -\sum_{i=0}^{L-1} p(z_i)\log_2 p(z_i)$ | Medida de aleatoriedad |

Tabla 3.1: Descriptores de textura

### Análisis multiresolución

La representación de textura de la imagen también puede darse mediante la implementación de la técnica de análisis multiresolución que consiste en descomponer una imagen mamográfica en otra con menor resolución, esto manteniendo una colección de coeficientes que son necesarios para recuperar la imagen original.

Existe una estructura para la representación de imágenes de distintas resoluciones presentada por Burt y Adelson (1983), llamada pirámide de imágenes. Dicha pirámide contiene una colección de imágenes de diferentes resoluciones ordenadas de tal manera que la base contiene una representación de alta resolución de la imagen procesada, mientras que cada nivel contiene una de menor resolución y así, la punta de la pirámide corresponde a la aproximación de menor resolución de la imagen, tal como se muestra en la Figura 3.9.

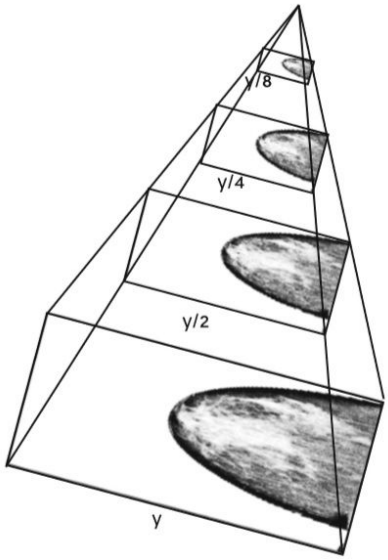


Figura 3.9: Pirámide de imágenes para el análisis multiresolución (Arce 2018)

Para llevar a cabo el análisis multiresolución se implementó la transformada wavelet discreta (DWT, por sus siglas en inglés) de Daubechies (1992). En los experimentos de Reyad (2014) fue la que mejor resultado proporcionó. La imagen mamográfica se descompone en términos de espacio y frecuencia por diversos coeficientes, de los cuales se denota como $A$ a una sub-banda de coeficientes de baja frecuencia, mientras que los coeficientes de alta frecuencia se dividen en tres sub-bandas: $V$, coeficientes de alta frecuencia vertical, $H$, coeficientes de alta frecuencia Horizontal y $D$, coeficientes de alta frecuencia diagonal. Con base a lo anterior, se estableció la DWT en su último nivel con la finalidad de convertir la sub-banda más baja en un vector de características (Ver Figura 3.10).

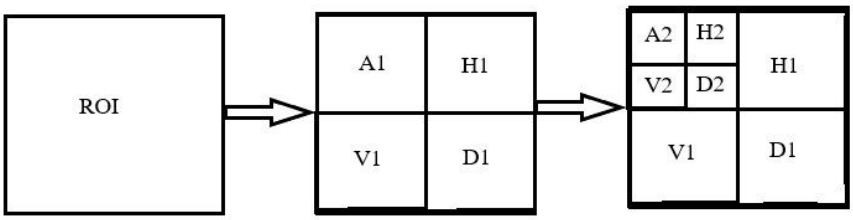


Figura 3.10: Descomposición de resolución con DWT

### Haralick

La extracción de características por el método de Haralick (1973) se realiza mediante un análisis de descriptores de textura que son un conjunto de medidas de textura basadas en una matriz de co-ocurrencia de distintos niveles de grises (GLCM, por sus siglas en inglés) que brindan información de la relación espacial entre los niveles de grises para coadyuvar en el análisis de textura. Las medidas estadísticas de textura de GLCM se muestran en la Tabla 3.2. Además, de la expresión correspondiente a cada una de las medidas, entre ellas el segundo momento angular, el contraste, correlación, varianza, momento de diferencia inversa, suma promedio, suma de varianza, suma de entropía, entropía, diferencia de varianza, diferencia de entropía y las medidas de información de la correlación. El coeficiente de correlación máximo, tal como lo menciona el mismo Haralick (1973) no ha sido calculado debido a una inestabilidad computacional.

| No. | **Medida estadística** | **Expresión** |
|---|---|---|
| **1** | Segundo momento angular | $f_1 = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} \left( \frac{P(i,j)}{R} \right)^2 = \sum_i \sum_j p(i,j)^2$ |
| **2** | Contraste | $f_2 = \sum_{k=0}^{N_g-1} k^2 \left\{ \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} \delta_{\lvert i-j \rvert, k} p(i,j) \right\}$ |
| **3** | Correlacion | $f_3 = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} (ij) p(i,j) - \mu_x \mu_y}{\sigma_x \sigma_y}$ |
| **4** | Varianza | $f_4 = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} (i-\mu)^2 p(i,j)$ |
| **5** | Momento de diferencia inversa | $f_5 = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} (i-\mu)^2 p(i,j)$ |
| **6** | Suma promedio | $f_6 = \sum_{i=2}^{2N_g} i p_{x+y}(i)$ |
| **7** | Suma de varianza | $f_7 = \sum_{i=2}^{2N_g} (i - f_8)^2 p_{x+y}(i)$ |
| **8** | Suma de entropía | $f_8 = -\sum_{i=2}^{2N_g} p_{x+y}(i) \log(p_{x+y}(i))$ |
| **9** | Entropía | $f_9 = -\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i,j) \log(p(i,j))$ |
| **10** | Diferencia de varianza | $f_{10} =$ variance of $p_{x-y}$ |
| **11** | Diferencia de entropía | $f_{11} = -\sum_{i=0}^{N_g-1} p_{x-y}(i) \log(p_{x-y}(i))$ |
| **12** | Medida de información de la correlación I | $f_{12} = \frac{f_9 - HXY1}{\max(HX, HY)}$ |
| **13** | Medida de información de la correlación II | $f_{13} = [1 - \exp(-2(HXY2 - f_9))]^{1/2}$ |

Tabla 3.2: Medidas estadísticas de textura por Haralick (1973)

## 3.2. Multiclasificación de los niveles de densidad mamográfica

La multiclasificación está desarrollada bajo un programa genético, mismo que está definido por un conjunto de individuos en forma de árbol, (ver Figura 3.11) que conforman una población, estos a su vez pueden reproducirse, cruzarse y mutarse en cada generación con la finalidad de preservar las mejores características, de tal manera que puedan mejorar los descendientes progresivamente en cada época. Por otra parte, se hace una implementación de la validación cruzada estratifcada para mantener un balance de carga entre la muestras por clase.

### 3.2.1. Programa genético

La multiclasificación con respecto al programa genético toma en cuenta que los individuos están evaluados como un árbol de regresión del cual se desprenden cuatro subárboles, donde cada uno de estos representa a un modelo clasificador para cada nivel de densidad mamográfica, tal como se muestra en la Figura 3.11. El objetivo de apreciar en forma de árbol a los individuos permite que al finalizar la evolución, además de generarse en una sola ejecución, se obtenga la solución desde el nodo raíz del mejor árbol, para después utilizarse en la multiclasificación. En otras palabras, cuando se finaliza la evolución de los individuos, estos tienen la capacidad de evaluarse en el nodo raíz para determinar un único modelo clasificador. Dentro de este punto, cada uno de los subárboles está compuesto por un conjunto de hojas que corresponden a un conjunto de terminales$T = \{IN_0, IN_1, IN_2, ...\}$, denotadas por números aleatorios entre 0 y 1, y que contienen las características extraídas de las imágenes mamográficas. Por otra parte, los nodos internos están compuestos por el conjunto de primitivas o funciones $F = \{+, -, \times, mydiv, mypower2, cos, sin, mysigmoid, switch\}$. Cabe mencionar que, la evaluación de cada árbol regresa un valor flotante que representa a cada una de las diferentes clasificaciones.

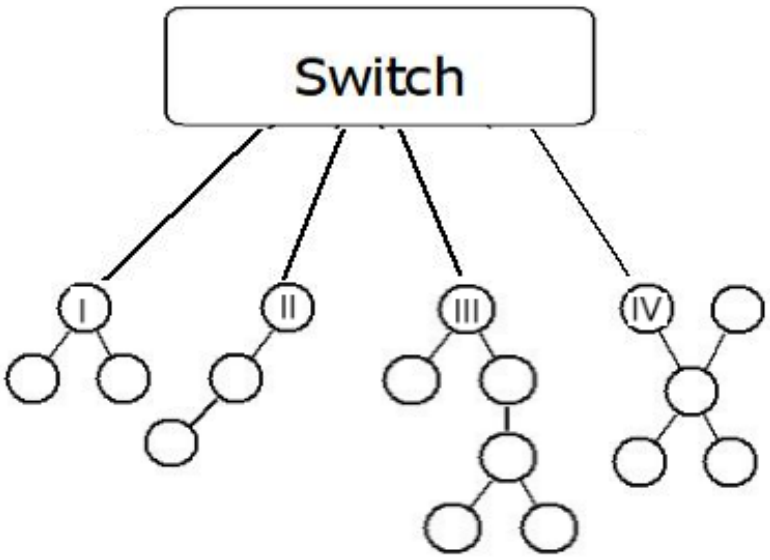


Figura 3.11: Representación del individuo clasificador

Bajo el marco de trabajo del programa genético, la multiclasificación se desarrolla en funcion de los siguientes pasos:

1. Se establecen los parámetros evolutivos los cuales corresponden a la tasa de cruzamiento $CXPB = 0.8$, mutación $MUTPB = 0.15$, el número de generaciones $NGEN = 100$, la frecuencia $FREQ = 10$ y número de Folds $n\_splits = 10$ que también se utilizan en la implementación de la paralelización con el modelo por islas que será explicado a detalle en la sección 3.3. También, teniendo en cuenta que el algoritmo implementado en este proyecto está basado en $(\mu + \lambda)$ que evalúa los individuos con un fitness inválido, es así que se requieren definir los parámetros evolutivos tales como, $\mu = 100$ que representa al número de individuos a seleccionar en cada generación y $\lambda = 200$ como el número de hijos que se producen en cada generación. Partiendo de $\lambda$, los descendientes son generados por una función de variación que ejecuta el cruzamiento, la mutación o la reproducción, y a su vez gestiona el número de generaciones definidas para finalmente regresar una tupla con la población final de la evolución. Lo anterior se sintetiza en la Tabla 3.3.

| | |
|---|---|
| Tamaño de la población | $\mu = 100$ |
| Probabilidad de cruzamiento | 0.98 |
| Probabilidad de mutación | 0.15 |
| Selección de descendientes | Basado en fitness $\mu + \lambda$ |
| Criterio de finalización | Cuando el algoritmo termina 100 generaciones |

Tabla 3.3: Configuración del programa genético

2. Se leen las características extraídas con base a la sección 3.1.2, propiedades que se almacenan en un archivo y se gestionan desde ahí para guardarse en una lista de listas, así mismo se hace la implementación de la validación cruzada estratificada que se desarrolla inmersa en el pograma genético. Lo anterior con la finalidad de hacer una distribución justa para cada nivel, debido a que existe una desproporción de casos dentro del conjunto de datos. Cada individuo es evaluado a través de una función de fitness, donde se genera una umbralización para la multiclasificación haciendo una selección de rango dinámico por clase bajo el esquema de Zhang y Smart (2004). Así mismo, se implementa la metodología de cruzamiento y mutación como operaciones genéticas para la mejora evolutiva del modelo clasificador.

A grandes rasgos, la funcionalidad del programa genético para la multiclasificación sigue una secuencia basada en la perspectiva de Koza (1992), con adecuaciones que brindan un mejor resultado para este problema en específico, de tal manera se muestra el Algoritmo 1 para la multiclasificación de los niveles de densidad mamográfica con programación genética:

**Algorithm 1** Algoritmo de multiclasificación de los niveles de densidad mamográfica con Programación Genética

```
Se genera una población inicial
Calcular la función de evaluación de cada individuo.
while not termine do
  /* producir nueva generación */
  for Cada subpoblación do
    Seleccionar individuos de la generación anterior.
    Cruzar los individuos obteniendo dos descendientes.
    Mutar los descendientes.
    Calcular la función de fitness para cada individuo.
    Selección de centro dinámico para cada individuo
    Reservar los individuos con el mejor fitness.
  end for
  if La población converge then
    Terminado := TRUE
  end if
end while
```

### Validación Cruzada Estratificada

En la validación cruzada, el conjunto de datos es aleatoriamente dividido en subconjuntos llamados en este entrorno como *k-Folds* mismos que son muatuamente excluyentes y en su mayoría del mismo tamaño. Para este trabajo se utilizan $k = 10$ Folds, siguiendo la apreciación de Markatou et al. (2005), los cuales serán entrenados y probados $k$ veces con el objetivo de obtener una mejor precisión al momento de clasificar, pecisión que está dada principalmente en la validación como la división del total de clasificaciones correctas por el número de entidades en el conjunto de datos. La implementación de la validación cruzada estratificada permite que durante el proceso de división de los datos para definir los conjuntos de prueba y entrenamiento, estos se reorganicen de tal manera que cada uno de los Folds represente un subconjunto de datos con al menos una característica de cada tipo, con la finalidad de evitar el sesgo de la información que bien podría darse en una validación cruzada regular. La Figura 3.12 muestra la distribución de los datos a través de la validación cruzada estratificada.

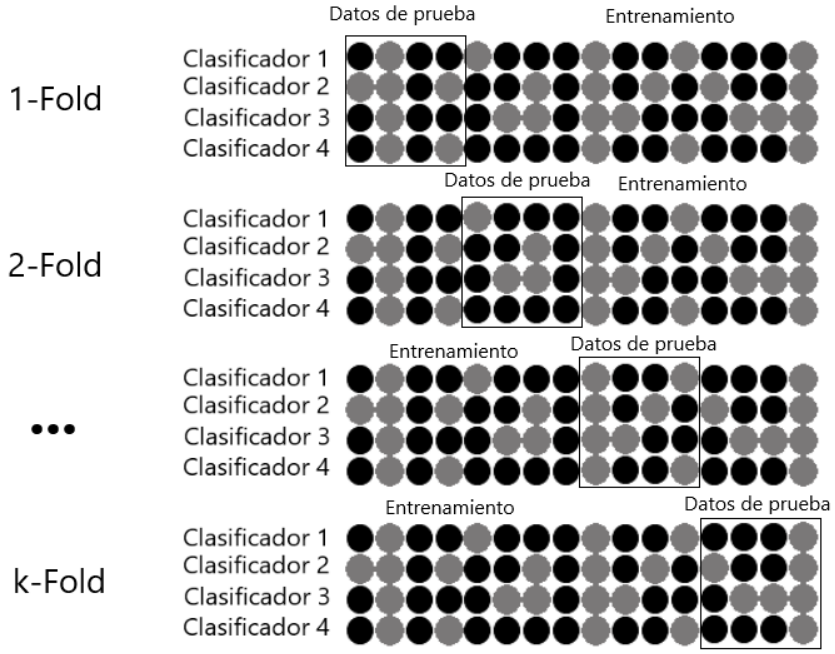


Figura 3.12: Esquema de la validación cruzada estratificada

### Función de evaluación

La evaluación se define como un problema de maximización, debido a que consiste en la suma de coincidencias, por lo tanto el individuo es un árbol a maximizar. Para la función de Fitness se transforma la expresión del árbol en una función que evalua si los datos de entrada coinciden con los de entrenamiento en cada nivel, y al coincidir se genera un incremento a la variable designada para el conteo de coincidencias de los individuos con base a la matriz de confusión, como se muestra en la Figura 3.13. De manera específica, la evaluación de las coincidencias se trabaja bajo el concepto de la medida de precisión *F1-Score* que está directamente ligada a dos métricas, tales como, la precisión y recall que son parte de las medidas estadísticas de rendimiento para la clasificación.

### Matriz de confusión

En apoyo a la implementación del *F1-Score*, es importante reconocer que en una matriz de confusión se hace una analogía de términos importantes para la clasificación, estos son: los Verdaderos Positivos (*TP*, por sus siglas en inglés) que genera una coincidencia cuando los datos de entrada corresponden con los de entrenamiento, los Verdaderos Negativos (*TN*, por sus siglas en inglés) que representa la no coincidencia entre los datos de entrada y entrenamiento, los Falsos Positivos (*FP*, por sus siglas en inglés) afirman una coincidencia entre los datos de entrada y de entrenamiento mientras que no existe esta, y los Falsos Negativos (*FN*, por sus siglas en inglés) que afirman que no se genera una coincidencia, mientras que si existe.

| | Predicción (+) | Predicción (-) |
|---|---|---|
| Objetivo (+) | Verdadero Positivo (TP) | Falso Negativo (FN) |
| Objetivo (-) | Falso Positivo (FP) | Verdadero Negativo (TN) |

Figura 3.13: Matriz de confusión

### Precisión y recall

Una vez expresada la estructura de la matriz de confusión, también la medida *F1-Score* requiere de las medidas de precisión y recall, tal como se mencionaba al inicio de esta sección. La *precisión* es una medida que permite determinar sí el costo de los *FP* es alto, tanto que interfiera en una correcta clasificación. Con ello se determina qué tan exacto es el modelo y si existe un número considerable de aquellos *TP*. La precisión está dada bajo la Ecuación 3.3

$$precision = \frac{TP}{TP + FP} \tag{3.3}$$

De igual forma, *recall* es una medida para calcular cuántos *TP* se han detectado y determinar sí el costo de los mismos tiene coherencia entre los *FN* obtenidos, es así que se puede asegurar que durante la evaluación no habrá un valor independiente respecto a los correcta-

mente clasificados y los correctamente no clasificados. Esta medida se puede calcular con la Ecuación 3.4.

$$Recall = \frac{TP}{TP + FN} \tag{3.4}$$

### F1-score

Con base a lo anteriormente definido, la función de fitness para obtener el mejor individuo clasificador está dada por la Ecuación 3.5, misma que se define como el doble del producto de la precision y el recall sobre la suma de las mismas medidas.

$$F1 - score = 2\,\frac{precision * recall}{precision + recall} \tag{3.5}$$

El uso de la medida F1-score representa la métrica más adecuada debido a la desproporción existente de clases, por lo que resulta aún mejor que la medida de exactitud, esto debido a que se implementa cuando se desea buscar un equilibrio entre los costos de clasificación dados por la precisión y el recall.

### Selección de rango

Como parte de la clasificación dentro del programa genético se realizó una umbralización que se determinó a través del cálculo dinámico de límites, donde cada uno de ellos se generan al calcular el centro del valor de salida del programa genético para cada clase, tal como se describe en el siguiente algoritmo:

**Algorithm 2** Algoritmo de selección de rango dinámico por clase

Se inicializan los centros con un valor estático

Para cada clase $C$, se calcula el centro de cada clase de acuerdo a la Ecuación 3.6.

$$Center_c = \frac{\sum_{p=1}^{M} \sum_{\mu_c=1}^{L} ProgOut_{p\mu c}}{MXL} \tag{3.6}$$

donde $M$ son los individuos de la población, $p$ el índice de población, $L$ el conjunto de entrenamiento por clase, $\mu_c$ el índice de clase y $ProgOut_{\mu c}$ es el valor flotante de salida por individuo del conjunto de entrenamiento por clase.

Se calcula el límite tomando el punto medio de dos centros de clases adyacentes.

Se hace la clasificación basada en los nuevos límites y la Ecuación 3.6.

### Cruzamiento

El cruzamiento entre dos individuos consiste en elegir aleatoriamente alguno de los cuatro subarboles, después se generará una iteración hasta encontrar la raíz de cada subarbol de ambos individuos y parte desde el subárbol encontrado para generar el cruzamiento tal como se muestra en las Figuras 3.14 y 3.15.

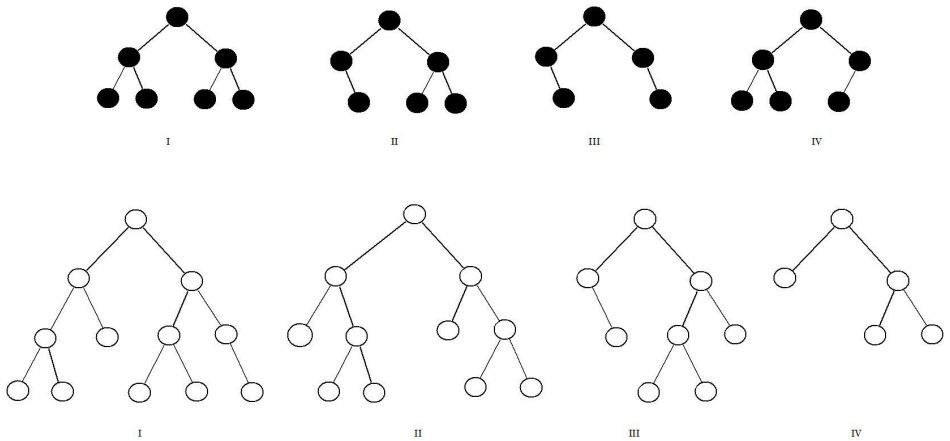


Figura 3.14: Subárboles antes del cruzamiento

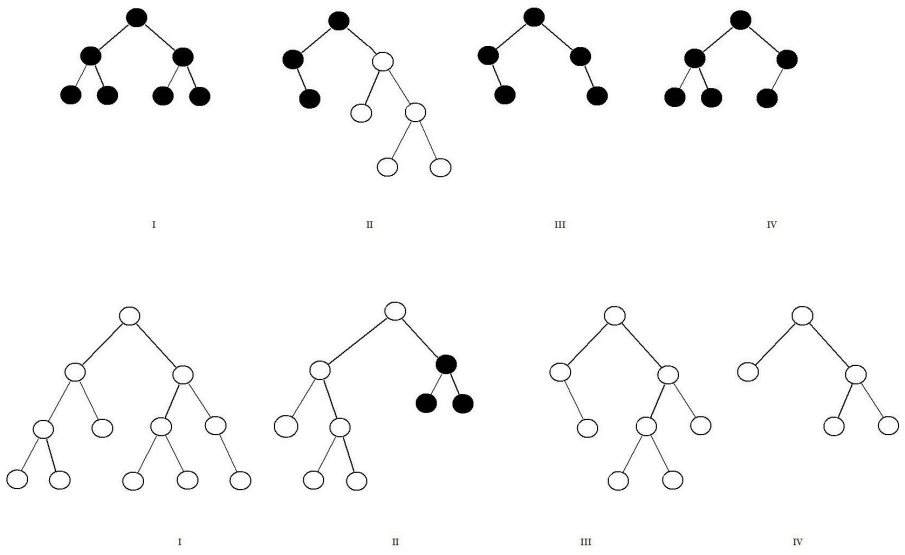


Figura 3.15: Subárboles después del cruzamiento

### Mutación

La mutación afecta a un individuo de la población, donde se selecciona de manera aleatoria el arbol (individuo) y se elige aleatoriamente un nodo del subárbol que será reemplazado para devolver un individuo ya mutado en forma de tupla. Una representación esquemática de lo mencionado se puede observar en las Figuras 3.16 y 3.17.

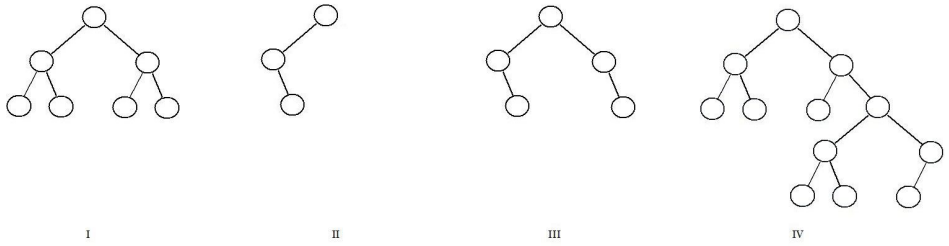


Figura 3.16: Subárboles antes de la mutación

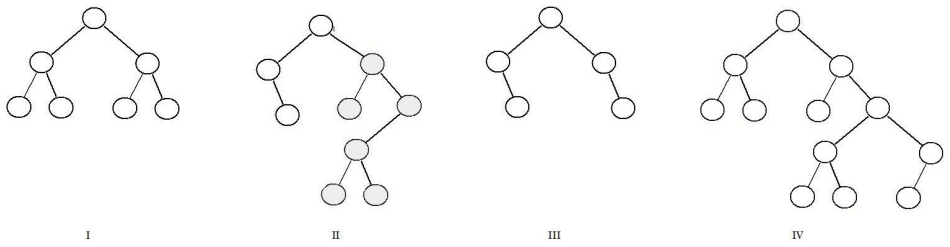


Figura 3.17: Subárboles después de la mutación

Finalmente, una vez definido el tamaño de la población, el número de generaciones, las probabilidades de mutación y cruzamiento, se hace la ejecución del programa genético para todos los individuos en espera del mejor clasificador. De forma experimental también se ejecutan otros clasificadores para generar una comparativa, entre ellos están Nearest Neighbors, Decision Tree, Random Forest, SVM, y Naive Bayes.

## 3.3. Paralelización

Como medida de reducción de tiempo de ejecución del método de multiclasificación se ha recurrido a la implementación del mismo en un ambiente paralelo. Ahora bien, el algoritmo implementado para la multiclasificación de los niveles de densidad mamográfica es apto para la paralelización. La idea principal es usar una operación de mapeo que aplique una función a cada elemento de una secuencia. En este programa genético un ejemplo es la función de evaluación (fitness), así, de manera predeterminada cada caja de herramientas (toolbox) se registra con la función de mapeo estándar para, dicha función de mapeo distribuye tareas concurrentes en diversos entornos que ejecutarán parte del programa de forma independiente y reorganizando la jerarquía de ejecución se sigue el modelo por islas que será explicado a continuación.

### 3.3.1. Modelo por islas

El paradigma del modelo de islas consiste en que múltiples poblaciones evolucionen por separado e intercambien individuos de manera continua en forma de migración. En esta inves-

tigación se aborda la paralelización mediante el modelo de islas paralelas (ver Figura 3.18), así como su rediseño, implementación y problemas asociados relacionados con el contexto de ejecución. Los resultados preliminares demuestran la efectividad de los enfoques propuestos y sus capacidades para explotar completamente la arquitectura. También se centran en el modelo cooperativo sincrónico de la isla, en el cual se despliegan simultáneamente para cooperar en el cálculo de soluciones mejores y más robustas. También, se genera una migración entre islas con el objetivo de retrasar la convergencia global, de manera esquemática se muestra como es la representación del modelo basado en islas siguiendo la topología de anillo en la Figura 3.19.

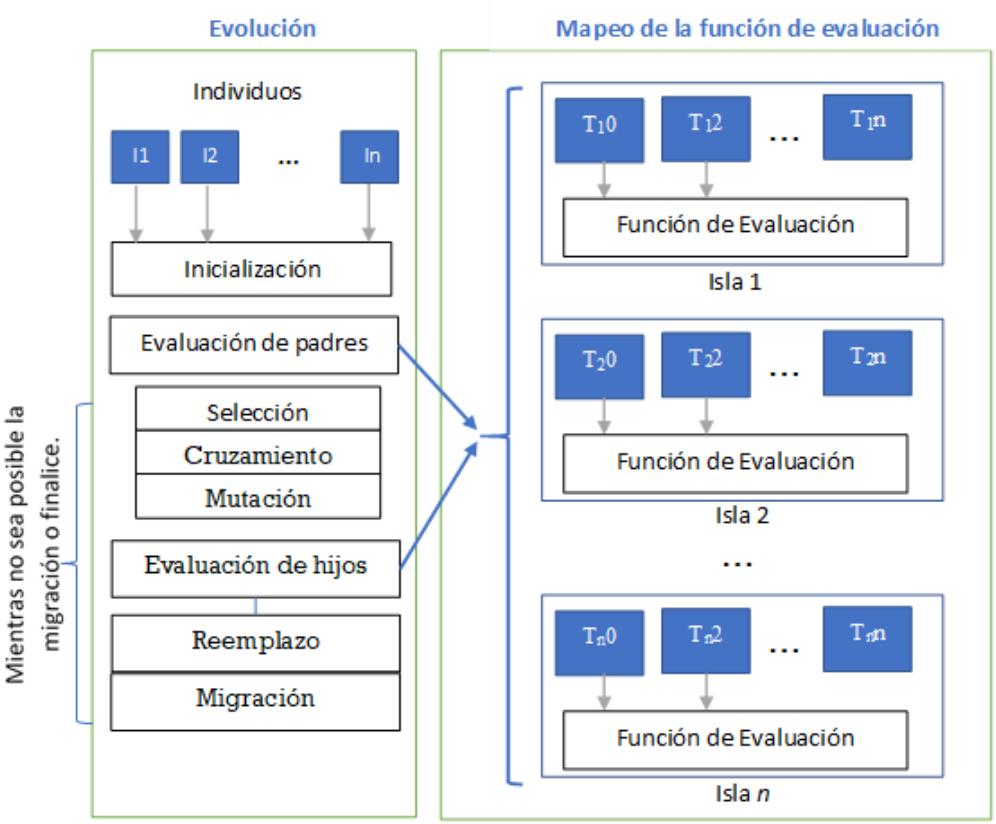


Figura 3.18: Paralelización implementado el modelo por islas

De acuerdo con la Figura 3.18, el punto de partida para la paralelización del método de multiclasificación de los niveles de densidad mamográfica es la inicialización de la evolución que será ejecutada a través de la CPU, mientras que la evaluación será llevada a cabo desde

la individualidad de las islas con apoyo de la población ya inicializada. La organización de cada worker depende de la ejecución simultanea del mismo programa dividida en 10 islas, que se traduce a un total de 9 workers y 1 root worker (master), misma que se hace de forma independiente en cada uno de los procesadores, por lo cual, un subproceso asociado a un individuo ejecuta la misma función de evaluación y finalmente los resultados de la evaluación serán procesados en conjunto al finalizar la migración, esta secuencia se puede visualizar en el Algoritmo 3.

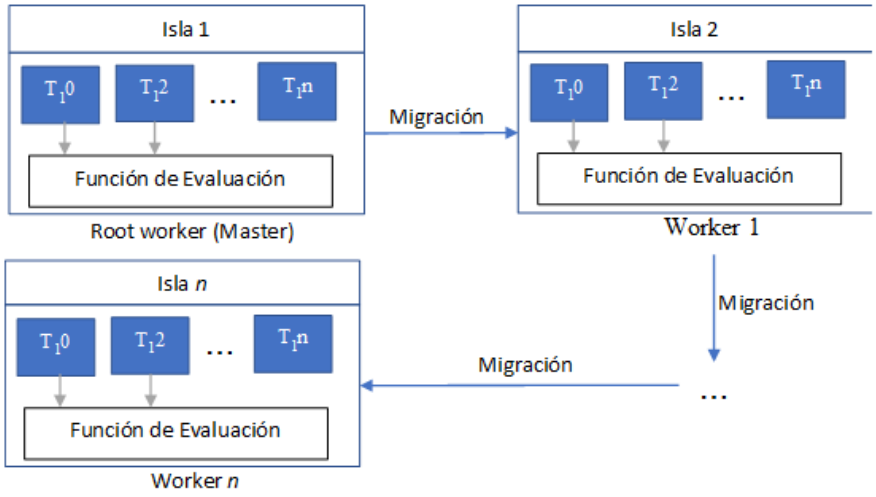


Figura 3.19: Esquematización del modelo por islas con topología de anillo

La topología de anillo dentro de la paralelización mantiene el orden de ejecución mediante la comunicación de cada una de las islas, condición favorable para la migración entre islas siguiendo la estructura topológica de anillo que brinda una comunicación circular y retrasa la convergencia global.

**Algorithm 3** Algoritmo GP basado en islas asíncronas

```
Crea una población aleatoria
while no termine do
  //en paralelo
  Evalua y selecciona a los individuos por su funcion fitness
  Se cruza y muta la población
  Selecciona los mejores n individuos y se envían al master
  Recibe un conjunto de n nuevos individuos desde el master
  Verifica que todos los workers terminen
  for Cada worker durante su ejecución do
    Reemplaza n los peores individuos de la población por n nuevos individuos recibidos
    del master
  end for
end while
Al mismo tiempo. El master ejecuta los siguientes pasos:
for Cada iteración donde hay comunicación do
  for Cada poblacion p do
    Recibe un conjunto de n individuos desde p
    Envía otros individuos de acuerdo con la topología de anillo.
  end for
end for
```

### 3.3.2. Tiempo de ejecución y speedup

Las métricas implementadas para el método de multiclasificación en paralelo son el tiempo de ejecución y la aceleración (o speedup), mismas que permiten demostrar la mejora obtenida al momento de paralelizar el método de multiclasificación.

Tomando como base que la ejecución se lleva a cabo en varios procesadores y puede que no todos ellos inicien o finalicen la ejecución al mismo tiempo, por esta razón se toma como tiempo de ejecución al que transcurre desde que empieza la ejecución del primer procesador, hasta que finaliza el último de los procesadores.

La aceleración de la ejecución tratada en este caso como el Speedup mide la velocidad que se gana al paralelizar el algoritmo de multiclasificación respecto al algoritmo secuencial de multiclasificación, para ello se implementó la Ecuación 3.7

$$S(n,p) = \frac{t(n)}{t(n,p)} \tag{3.7}$$

donde $t(n)$ es el tiempo de ejecución secuencial del método de multiclasificación, y $t(n,p)$ el tiempo de ejecución del método de multiclasificación en paralelo.

# CAPÍTULO 4

# Resultados

**Contenido**



Como prueba del trabajo desarrollado es fundamental la representación de los resultados, los cuales se expresan con base a la multiclasificación. Por otra parte, se consideran también los resultados del método de multiclasificación en un ambiente paralelo, esto a través de una comparativa con el coste de tiempo del método de multiclasificación en una ejecución secuencial.

## 4.1. Consideraciones

Las características de la computadora en la cual se realizaron los experimentos es una computadora Dell con memoria de 15 GB, procesador Intel Core i7-4770K de 3.5 GHz con 8 cores, disco duro de 126 GB y un sistema Base Ubuntu 16.04.06 LTS 64 bit. El entorno de trabajo sobre el programa genético se desarrolló bajo el lenguaje Python 3.7.4, con la biblioteca DEAP 1.3.0 para la implementación del programa genético.

Las imágenes fueron tomadas de la base de datos de INBreast con extensión DICOM (Digital Imaging and Communication On Medicine), y la base de datos MIAS con extensión PGM (Portable Gray Map Image). Ambas son imágenes en escala de grises. Las imágenes desde el pre-procesamiento adquieren modificaciones, tal como se describió anteriormente en

el capítulo 3 mismas que enmarcan una Región de Interésc con la que se trabaja para obtener los resultados posteriores.

## 4.2. Resultados de la multiclasificación

A lo largo de esta sección se podrán apreciar los resultados de la multiclasificación de los niveles de densidad mamográfica con Programación Genética para los conjuntos de datos MIAS e INbreast.

### 4.2.1. Resultados con el conjunto de datos INbreast

A continuación, se muestra en formato de tabla el desempeño del clasificador multiclase con programación genética en comparativa con el desempeño de otros clasificadores, esto para el conjunto de datos INBreast. Cada tabla expone el resultado con base a las medidas de F1-score, precision (presición), accuracy (exactitud), recall (sensibilidad) y confusion matrix (matriz de confusión), lo anterior bajo un esquema de validación cruzada estratificada de 10-fold.

Los resultados correspondientes al clasificador Nearest Neighbors sobre la base de datos INbreast se puede visualizar en la Tabla 4.1.

| Nearest Neighbors | | | | | |
|---|---|---|---|---|---|
| **Fold** | **F1-score** | **Precision** | **Accuracy** | **Recall** | **Confusion matrix** |
| 1 | 0.18866429269655075 | 0.17637721755368816 | 0.20952380952380953 | 0.20952380952380953 | [6 6 2 0]<br>[6 6 3 0]<br>[3 5 1 1]<br>[2 1 0 0] |
| **2** | **0.25329040959040954** | **0.25262539682539686** | **0.26317142857142855** | **0.26317142857142855** | [3 7 2 1]<br>[ 5 5 3 1]<br>[3 5 1 0]<br>[0 1 0 1] |
| 3 | 0.24964912280701755 | 0.2439153439153439 | 0.25095238095238093 | 0.25095238095238093 | [8 4 2 0]<br>[6 6 3 0]<br>[4 4 2 0]<br>[0 1 2 0] |
| 4 | 0.17205882352941174 | 0.14269005847953212 | 0.20952380952380953 | 0.20952380952380953 | [6 6 2 0]<br>[6 7 1 1]<br>[4 5 0 1]<br>[2 1 0 0] |
| 5 | 0.2400357083115704 | 0.2459018759018759 | 0.2557142857142857 | 0.2557142857142857 | [3 8 1 2]<br>[6 8 1 0]<br>[4 5 1 0]<br>[2 1 0 0] |
| 6 | 0.20617852875917392 | 0.20019257703081234 | 0.21428571428571427 | 0.21428571428571427 | [3 6 5 0]<br>[7 5 2 1]<br>[4 5 1 0]<br>[3 0 0 0] |
| 7 | 0.24881858210062834 | 0.24708478513356556 | 0.25585365853658536 | 0.25585365853658536 | [7 4 1 2]<br>[4 6 4 0]<br>[3 4 2 1]<br>[1 2 0 0] |
| 8 | 0.172222222222222 | 0.1955357142857143 | 0.2 | 0.2 | [6 2 4 1]<br>[5 5 3 1]<br>[2 5 1 2]<br>[1 2 0 0] |
| 9 | 0.2106739846935926 | 0.1952787952787953 | 0.2333333333333333 | 0.2333333333333333 | [5 6 2 0]<br>[4 7 3 0]<br>[4 5 1 0]<br>[1 0 1 0] |
| 10 | 0.20293069839349331 | 0.19586466165413535 | 0.21052631578947367 | 0.21052631578947367 | [5 5 1 2]<br>[6 3 1 4]<br>[3 5 0 1]<br>[0 2 0 0] |

Tabla 4.1: Resultados de la multiclasificación con Nearest Neighbors del conjunto de datos INbreast

La Tabla 4.1 muestra que para la función F1-score se obtuvo un máximo del 25 % respecto a los folds, manteniendo el mismo porcentaje de precisión y un 26 % de exactitud y recall, así mismo, para la matriz de confusión se logra percibir que existe un mayor valor de casos correctamente clasificados.

Los resultados del clasificador Decision Tree se muestran directamente en la Tabla 4.2.

| Decision Tree | | | | | |
|---|---|---|---|---|---|
| **Fold** | **F1-score** | **Precision** | **Accuracy** | **Recall** | **Confusion matrix** |
| 1 | 0.25629078982020156 | 0.22557477025898079 | 0.27619047619047616 | 0.27619047619047616 | [7 3 4 0]<br>[ 4 10 0 1]<br>[1 6 2 1]<br>[0 0 2 1] |
| 2 | 0.25015725275192314 | 0.269212962962963 | 0.2523809523809524 | 0.2523809523809524 | [5 6 2 1]<br>[2 7 5 1]<br>[1 3 5 1]<br>[1 0 0 2] |
| 3 | 0.229139515455305 | 0.2262962962962963 | 0.2357142857142857 | 0.2357142857142857 | [4 8 1 1]<br>[9 3 3 0]<br>[3 3 4 0]<br>[0 1 1 1] |
| 4 | 0.2113275613275613 | 0.24548229548229553 | 0.2857142857142857 | 0.2857142857142857 | [4 3 5 2]<br>[0 7 4 4]<br>[5 3 1 1]<br>[1 0 2 0] |
| **5** | **0.2648435127674258** | **0.2286522301228184** | **0.3** | **0.3** | [10 3 1 0]<br>[0 7 8 0]<br>[0 4 4 2]<br>[0 3 0 0] |
| 6 | 0.20198412698412697 | 0.2248344792462439 | 0.2857142857142857 | 0.2857142857142857 | [5 5 2 2]<br>[4 4 3 4]<br>[2 5 3 0]<br>[0 3 0 0] |
| 7 | 0.19273810837655884 | 0.2089430894308943 | 0.1926829268292683 | 0.1926829268292683 | [5 8 1 0]<br>[4 5 3 2]<br>[2 5 2 1]<br>[1 2 0 0] |
| 8 | 0.25149243560038157 | 0.26000000000000004 | 0.35 | 0.35 | [5 8 0 0]<br>[4 4 6 0]<br>[1 1 5 3]<br>[0 2 1 0] |
| 9 | 0.2169466548776894 | 0.188034188034188 | 0.2076923076923077 | 0.2076923076923077 | [3 8 2 0]<br>[1 5 7 1]<br>[1 2 4 3]<br>[0 0 2 0] |
| 10 | 0.2263729977116705 | 0.2345205118058303 | 0.24210526315789475 | 0.24210526315789475 | [8 2 2 1]<br>[8 3 2 1]<br>[3 3 2 1]<br>[0 1 1 0] |

Tabla 4.2: Resultados de la multiclasificación con Decision Tree del conjunto de datos INbreast

Siguiendo el valor representativo de las funciones aplicadas para la obtención de resultados resultados, se logra ver como el mejor caso para F1-score representa un 26 %, para precisión

mantiene un valor de 22 %, mientras que para exactitud y recall un 30 %. Por otra parte, la matriz de confusión expresa que hay un mayor índice de casos correctamente clasificados.

Por otra parte, los resultados obtenidos al aplicar el clasificador Random Forest se expresan en la Tabla 4.3.

| Random Forest | | | | | |
|---|---|---|---|---|---|
| **Fold** | **F1-score** | **Precision** | **Accuracy** | **Recall** | **Confusion matrix** |
| 1 | 0.4099742599742599 | 0.396031746031746 | 0.4176190476190477 | 0.4176190476190477 | [9 2 2 1]<br>[ 3 9 1 2]<br>[1 7 0 2]<br>[0 0 0 3] |
| 2 | 0.37045570916538656 | 0.36519607843137253 | 0.38095238095238093 | 0.38095238095238093 | [8 5 1 0]<br>[7 5 3 0]<br>[2 3 3 2]<br>[0 2 1 0] |
| 3 | 0.3802437641723357 | 0.36313686313686316 | 0.40476190476190477 | 0.40476190476190477 | [9 5 0 0]<br>[7 4 4 0]<br>[2 4 4 0]<br>[0 0 3 0] |
| 4 | 0.3422318422318423 | 0.33664021164021163 | 0.35714285714285715 | 0.35714285714285715 | [8 4 2 0]<br>[4 5 6 0]<br>[5 3 2 0]<br>[1 0 2 0] |
| 5 | 0.4224362519201227 | 0.41456582633053224 | 0.4223809523809524 | 0.4223809523809524 | [9 5 0 0]<br>[5 5 5 0]<br>[2 3 5 0]<br>[1 2 0 0] |
| 6 | 0.41953495376645285 | 0.41725966702470463 | 0.42619047619047616 | 0.42619047619047616 | [9 1 2 2]<br>[6 7 2 0]<br>[2 6 2 0]<br>[0 2 1 0] |
| 7 | 0.4088353769200125 | 0.40433604336043355 | 0.4146341463414634 | 0.4146341463414634 | [7 5 2 0]<br>[4 6 2 2]<br>[3 3 4 0]<br>[1 1 1 0] |
| 8 | 0.419162685140946 | 0.4161538461538462 | 0.5 | 0.5 | [9 2 2 0]<br>[4 6 4 0]<br>[2 3 5 0]<br>[0 1 2 0] |
| 9 | 0.4060284543043163 | 0.404040404040404 | 0.4128205128205128 | 0.4128205128205128 | [6 4 2 1]<br>[2 9 3 0]<br>[3 2 5 0]<br>[0 0 2 0] |
| **10** | **0.4225864143116412** | **0.462159649122807** | **0.4473315789473685** | **0.4473315789473685** | [10 2 1 0]<br>[8 5 0 0]<br>[7 7 0 0]<br>[0 1 1 0] |

Tabla 4.3: Resultados de la multiclasificación con Random Forest del conjunto de datos INbreast

El valor máximo de F1-score es de 42 %, para la precisión incrementó el valor a 46 % , 44 % para la exactitud y el recall. Dentro de la matriz de confusión existe un valor mayor sobre los correctamente clasificados a diferencia de los incorrectamente clasificados.

Para SVM, siendo este el clasificador de menor rendimiento se aprecian los resultados en la Tabla 4.4.

| SVM | | | | | |
|---|---|---|---|---|---|
| **Fold** | **F1-score** | **Precision** | **Accuracy** | **Recall** | **Confusion matrix** |
| 1 | 0.18796992481203006 | 0.12755102040816327 | 0.35714285714285715 | 0.35714285714285715 | [ 0 14 0 0]<br>[ 0 15 0 0]<br>[ 0 10 0 0]<br>[0 3 0 0] |
| 2 | 0.18796992481203006 | 0.12755102040816327 | 0.35714285714285715 | 0.35714285714285715 | [ 0 14 0 0]<br>[ 0 15 0 0]<br>[ 0 10 0 0]<br>[0 3 0 0] |
| 3 | 0.18796992481203006 | 0.12755102040816327 | 0.35714285714285715 | 0.35714285714285715 | [ 0 14 0 0]<br>[7 4 4 0]<br>[2 4 4 0]<br>[0 0 3 0] |
| 4 | 0.18796992481203006 | 0.12755102040816327 | 0.35714285714285715 | 0.35714285714285715 | [ 0 14 0 0]<br>[ 0 15 0 0]<br>[ 0 10 0 0]<br>[0 3 0 0] |
| 5 | 0.18796992481203006 | 0.12755102040816327 | 0.35714285714285715 | 0.35714285714285715 | [ 0 14 0 0]<br>[ 0 15 0 0]<br>[ 0 10 0 0]<br>[0 3 0 0] |
| 6 | 0.18796992481203006 | 0.12755102040816327 | 0.35714285714285715 | 0.35714285714285715 | [ 0 14 0 0]<br>[ 0 15 0 0]<br>[ 0 10 0 0]<br>[0 3 0 0] |
| 7 | 0.1738359201773836 | 0.11659726353361097 | 0.34146341463414637 | 0.34146341463414637 | [ 0 14 0 0]<br>[ 0 14 0 0]<br>[ 0 10 0 0]<br>[0 3 0 0] |
| 8 | 0.18148148148148147 | 0.12249999999999998 | 0.35 | 0.35 | [ 0 13 0 0]<br>[4 6 4 0]<br>[2 3 5 0]<br>[0 1 2 0] |
| 9 | 0.18964683115626513 | 0.12886259040105194 | 0.358974358974359 | 0.358974358974359 | [ 0 13 0 0]<br>[ 0 14 0 0]<br>[ 0 10 0 0]<br>[0 2 0 0] |
| **10** | **0.19838056680161942** | **0.13573407202216067** | **0.3684210526315789** | **0.3684210526315789** | [ 0 13 0 0]<br>[ 0 14 0 0]<br>[0 9 0 0]<br>[0 2 0 0] |

Tabla 4.4: Resultados de la multiclasificación con SVM del conjunto de datos INbreast

Se alcanzó un máximo de 19 % del valor de F1-score, visualizando sus valores es posible detectar que en todos los folds correspondientes los resultados no varían de forma drástica.

Así mismo, hubo un 13 % de precisión y un 36 % de exactitud y recall. Para este caso, es posible reconocer que en la matriz de confusión hay un mayor índice de incorrectamente clasificados.

Para el clasificador Naive Bayes, se presenta en la Tabla 4.5 el conjunto de resultados.

| Naive Bayes | | | | | |
|---|---|---|---|---|---|
| **Fold** | **F1-score** | **Precision** | **Accuracy** | **Recall** | **Confusion matrix** |
| 1 | 0.3574074074074738 | 0.35145502645502645 | 0.36666666666666666 | 0.36666666666666666 | [4 7 2 1]<br>[9 2 4 0]<br>[3 6 1 0]<br>[2 0 1 0] |
| 2 | 0.3508634222919937 | 0.342063492063492 | 0.40476190476190477 | 0.40476190476190477 | [4 8 2 0]<br>[ 2 12 1 0]<br>[4 4 1 1]<br>[2 1 0 0] |
| 3 | 0.3463642213642214 | 0.3286435786435786 | 0.38095238095238093 | 0.38095238095238093 | [6 7 1 0]<br>[5 9 1 0]<br>[3 6 1 0]<br>[0 0 3 0] |
| 4 | 0.37127620231068506 | 0.3815295815295815 | 0.3857142857142857 | 0.3857142857142857 | [ 4 10 0 0]<br>[7 6 2 0]<br>[3 5 2 0]<br>[1 1 1 0] |
| **5** | **0.4667257816257816** | **0.48798730158730157** | **0.52636190476190477** | **0.52636190476190477** | [13 0 0 0]<br>[7 5 2 0]<br>[2 5 2 0]<br>[2 0 0 0] |
| 6 | 0.3980927118858153 | 0.40952380952380953 | 0.3333333333333333 | 0.3333333333333333 | [5 8 1 0]<br>[7 8 0 0]<br>[2 7 1 0]<br>[1 1 1 0] |
| 7 | 0.35609756097560976 | 0.32195121951219513 | 0.2926829268292683 | 0.2926829268292683 | [ 3 11 0 0]<br>[5 8 1 0]<br>[1 7 1 1]<br>[1 2 0 0] |
| 8 | 0.36085332909582076 | 0.3672794117647059 | 0.375 | 0.375 | [4 6 2 1]<br>[6 6 1 1]<br>[4 4 1 1]<br>[2 1 0 0] |
| 9 | 0.3743246719835348 | 0.44234056541748856 | 0.41025641025641024 | 0.41025641025641024 | [4 9 0 0]<br>[ 4 10 0 0]<br>[2 6 2 0]<br>[0 1 1 0] |
| 10 | 0.3393901420217209 | 0.3505720823798627 | 0.3684210526315789 | 0.3684210526315789 | [8 3 2 0]<br>[7 5 2 0]<br>[6 2 1 0]<br>[2 0 0 0] |

Tabla 4.5: Resultados de la multiclasificación con Naive Bayes del conjunto de datos INbreast

El valor máximo del F1-score es de 46 %, aseverando que para el problema de clasificación de los niveles de densidad mamográfica representa el mejor valor, de igual forma, en precisión alcanza el 48 %, en exactitud y recall se obtuvo un 52 %.

En la Tabla 4.6 se presentan los resultados obtenidos a través de la ejecución del programa genético.

| Our_GP | | | | | |
|---|---|---|---|---|---|
| **Fold** | **F1-score** | **Precision** | **Accuracy** | **Recall** | **Confusion matrix** |
| 1 | 0.41544771018455225 | 0.4328732459167242 | 0.4333333333333333 | 0.4333333333333333 | [4 5 4 1]<br>[2 9 3 1]<br>[0 9 1 0]<br>[2 0 1 0] |
| 2 | 0.4186429276654841 | 0.41206220679904895 | 0.4333333333333333 | 0.4333333333333333 | [7 3 4 0]<br>[9 4 2 0]<br>[2 4 3 1]<br>[1 2 0 0] |
| 3 | 0.4061473127262601 | 0.4101851851851852 | 0.40952380952380953 | 0.40952380952380953 | [6 4 3 1]<br>[7 5 3 0]<br>[5 3 1 1]<br>[0 0 2 1] |
| 4 | 0.39831583268516115 | 0.36331763474620615 | 0.3857142857142857 | 0.3857142857142857 | [2 5 6 1]<br>[2 6 0 7]<br>[1 2 3 4]<br>[0 1 1 1] |
| 5 | 0.43524544118917987 | 0.44024943310657596 | 0.4333333333333333 | 0.4333333333333333 | [5 3 3 3]<br>[6 5 3 1]<br>[2 5 3 0]<br>[1 1 0 1] |
| 6 | 0.3887092731829574 | 0.39052951879038835 | 0.30952380952380953 | 0.30952380952380953 | [3 6 5 0]<br>[7 8 0 0]<br>[1 7 1 1]<br>[0 2 0 1] |
| 7 | 0.32697665573146963 | 0.30465916197623518 | 0.3682926829268293 | 0.3682926829268293 | [ 4 10 0 0]<br>[5 7 2 0]<br>[4 4 0 2]<br>[0 3 0 0] |
| **8** | **0.45997407407407414** | **0.4473344322344322** | **0.4736** | **0.4736** | [7 5 1 0]<br>[5 6 3 0]<br>[1 3 5 0]<br>[1 0 1 0] |
| 9 | 0.30590520590520592 | 0.28394648829431437 | 0.3564102564102564 | 0.3564102564102564 | [0 4 8 1]<br>[0 5 9 0]<br>[0 4 5 1]<br>[0 1 1 0] |
| 10 | 0.3545051433136922 | 0.35471823498139284 | 0.3631578947368421 | 0.3631578947368421 | [5 2 6 0]<br>[10 3 0 1]<br>[2 5 2 0]<br>[1 1 0 0] |

Tabla 4.6: Resultados de la multiclasificación con el programa genético del conjunto de datos INbreast

Se puede observar que el mejor fold representa un 45 % de F1-score, un 48 % en precisión, en exactitud y recall mantiene un 52 % y los valores de la matriz de confusión demuestran que hay más casos correctamente clasificados que aquellos incorrectamente clasificados.

Así mismo, cabe mencionar que es un resultado inicial sobre este conjunto de datos (INbreast) sometido a una clasificación multiclase.

En la Figura 4.1, se puede ver el rendimiento del GP aquí propuesto (Our_GP) con base a la información contenida en las Tablas 4.1 a 4.6, especificando que Our_GP está entre los mejores clasificadores.

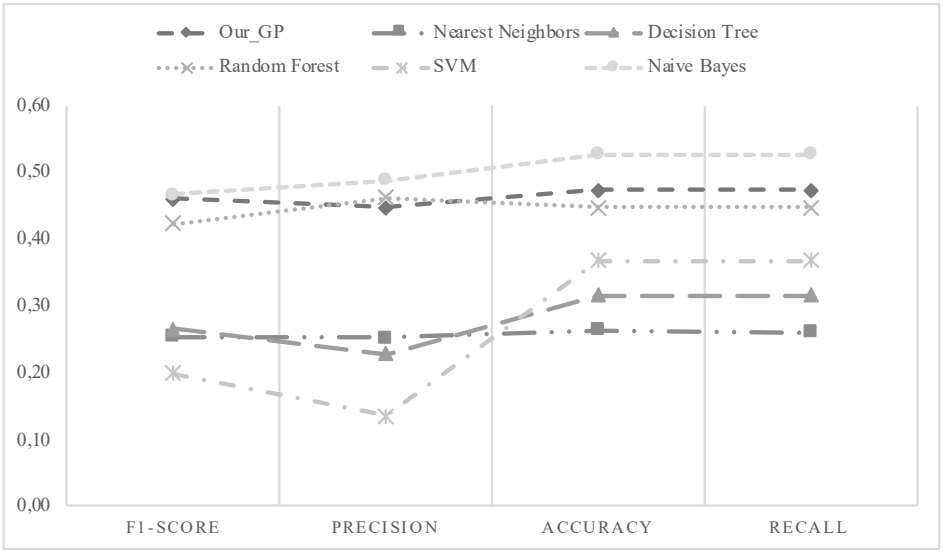


Figura 4.1: Gráfica de resultados de la multiclasificación con INbreast para todos los clasificadores

Es posible ver que, en el mejor de los casos de validación cruzada, Our_GP está solo por debajo de Naive Bayes, pero está justo por encima del resto de los clasificadores, una condición que lo hace competente para resolver el problema de multiclasificación de los niveles de densidad mamográfica.

### 4.2.2. Resultados con el conjunto de datos MIAS

Posteriormente, se muestra en formato de tabla el desempeño del clasificador multiclase con programación genética en comparativa con el desempeño de otros clasificadores, esto para el conjunto de datos MIAS. Cada tabla expone el resultado con base a las medidas de F1-score, precision, accuracy, recall y confusion matrix, lo anterior bajo un esquema de validación cruzada estratificada de 10-fold.

Los resultados para el clasificador Nearest Neighbors se expresan en la Tabla 4.7.

| Nearest Neighbors | | | | | |
|---|---|---|---|---|---|
| **Fold** | **F1-score** | **Precision** | **Accuracy** | **Recall** | **Confusion matrix** |
| 1 | 0.4267351384998443 | 0.4558823529411764 | 0.4294117647058824 | 0.4294117647058824 | [8 2 1]<br>[6 4 1]<br>[2 4 6] |
| 2 | 0.4238813474107592 | 0.4238813474107592 | 0.4294117647058824 | 0.4294117647058824 | [7 3 1]<br>[3 4 4]<br>[3 2 7] |
| **3** | **0.47795414462081126** | **0.5047619047619047** | **0.48484848484848486** | **0.48484848484848486** | [3 7 1]<br>[2 8 1]<br>[5 1 5] |
| 4 | 0.4359244620114186 | 0.4373737373737374 | 0.4363636363636364 | 0.4363636363636364 | [7 2 2]<br>[3 6 2]<br>[0 3 8] |
| 5 | 0.3947916666666667 | 0.3962606837606838 | 0.4 | 0.4 | [5 5 1]<br>[3 3 4]<br>[1 2 8] |
| 6 | 0.4650735294117647 | 0.49479166666666663 | 0.46875 | 0.46875 | [3 7 1]<br>[2 5 3]<br>[1 3 7] |
| 7 | 0.351627666642061 | 0.3570443893024538 | 0.3548387096774194 | 0.3548387096774194 | [5 5 0]<br>[5 1 4]<br>[3 3 5] |
| 8 | 0.4399228012131237 | 0.4564516129032258 | 0.4451612903225806 | 0.4451612903225806 | [6 1 3]<br>[2 5 3]<br>[0 2 9] |
| 9 | 0.36683644595359366 | 0.3742959549411163 | 0.3870967741935484 | 0.3870967741935484 | [7 2 1]<br>[5 2 3]<br>[3 5 3] |
| 10 | 0.4682438916365246 | 0.4741315136476427 | 0.4806451612903226 | 0.4806451612903226 | [9 1 0]<br>[3 4 3]<br>[1 5 5] |

Tabla 4.7: Resultados de la multiclasificación con Nearest Neighbors del conjunto de datos MIAS

Para la función F1-score se obtuvo un máximo del 47 %, una precisión de 50 %, en exactitud y recall el valor alcanzó un 48 %. Así mismo, es posible visualizar que los resultados correspondientes a la matriz de confusión arrojan un mayor índice de casos correctamente clasificados.

En la Tabla 4.8, se muestran los resultados de la clasificación con Decision Tree.

| Decision Tree | | | | | |
|---|---|---|---|---|---|
| **Fold** | **F1-score** | **Precision** | **Accuracy** | **Recall** | **Confusion matrix** |
| 1 | 0.5050980392156861 | 0.5061085972850678 | 0.5058823529411765 | 0.5058823529411765 | [8 2 1]<br>[1 7 3]<br>[2 1 9] |
| 2 | 0.5464705882352942 | 0.5464705882352942 | 0.5464705882352942 | 0.5464705882352942 | [8 2 1]<br>[1 7 3]<br>[2 2 8] |
| **3** | **0.5555555555555556** | **0.553113553113553** | **0.5757575757575758** | **0.5757575757575758** | [9 1 1]<br>[3 3 5]<br>[1 3 7] |
| 4 | 0.5284300494826811 | 0.5416666666666666 | 0.5363636363636364 | 0.5363636363636364 | [7 2 2]<br>[2 5 4]<br>[1 1 9] |
| 5 | 0.40544721177944854 | 0.4079059829059829 | 0.40625 | 0.40625 | [6 3 2]<br>[2 2 6]<br>[2 4 5] |
| 6 | 0.49829545454545454 | 0.5017361111111112 | 0.5 | 0.5 | [7 2 2]<br>[2 5 3]<br>[2 5 4] |
| 7 | 0.5075268817204302 | 0.5344086021505375 | 0.5129032258064516 | 0.5129032258064516 | [7 1 2]<br>[5 4 1]<br>[2 1 8] |
| 8 | 0.5171072843398818 | 0.5233300749429781 | 0.5129032258064516 | 0.5129032258064516 | [7 2 1]<br>[2 5 3]<br>[0 4 7] |
| 9 | 0.508886780518659 | 0.5274338564661145 | 0.5096774193548387 | 0.5096774193548387 | [6 3 1]<br>[0 7 3]<br>[1 1 9] |
| 10 | 0.5459324263738524 | 0.5444379276637341 | 0.5541935483870968 | 0.5541935483870968 | [10 0 0]<br>[0 8 2]<br>[1 4 6] |

Tabla 4.8: Resultados de la multiclasificación con Decision Tree del conjunto de datos MIAS

Decisión Tree, es el segundo mejor clasificador para éste problema despues de Our_GP, que representa un valor máximo para F1-score y precisión de 55 %, mientras que en exactitud y recall alcanzó un 57 %. Para los resultados de la matriz de confusión los valores sobre los casos correctamente clasificados son mayores a aquellos incorrectamente clasificados.

Los resultados del clasificador Random Forest se presentan en la Tabla 4.9.

| Random Forest | | | | | |
|---|---|---|---|---|---|
| **Fold** | **F1-score** | **Precision** | **Accuracy** | **Recall** | **Confusion matrix** |
| 1 | 0.5490835464620631 | 0.5644042232277527 | 0.5470588235294118 | 0.5470588235294118 | [8 3 0]<br>[2 7 2]<br>[2 3 7] |
| 2 | 0.5287270463741051 | 0.5580882352941176 | 0.5352941176470589 | 0.5352941176470589 | [9 1 1]<br>[ 0 10 1]<br>[1 5 6] |
| 3 | 0.5257575757575758 | 0.5324527324527325 | 0.5272727272727273 | 0.527272727272727 | [8 1 2]<br>[2 7 2]<br>[1 1 9] |
| **4** | **0.6282828282828283** | **0.626003626003626** | **0.6363636363636364** | **0.6363636363636364** | [9 2 0]<br>[2 4 5]<br>[0 3 8] |
| 5 | 0.5642986542443064 | 0.5717329545454546 | 0.5625 | 0.5625 | [6 4 1]<br>[1 5 4]<br>[2 2 7] |
| 6 | 0.5997072440087146 | 0.681781045751634 | 0.59375 | 0.59375 | [6 3 2]<br>[0 8 2]<br>[0 6 5] |
| 7 | 0.613968699214141 | 0.6139950372208436 | 0.6374193548387096 | 0.6374193548387096 | [8 2 0]<br>[3 5 2]<br>[2 1 8] |
| 8 | 0.6158256440540341 | 0.6387455197132617 | 0.6451612903225806 | 0.6451612903225806 | [7 2 1]<br>[2 5 3]<br>[0 3 8] |
| 9 | 0.6172787540699256 | 0.6287813620071686 | 0.6419354838709677 | 0.6419354838709677 | [8 2 0]<br>[1 7 2]<br>[0 3 8] |
| 10 | 0.6062755181549011 | 0.6081389578163771 | 0.6096774193548387 | 0.6096774193548387 | [10 0 0]<br>[2 5 3]<br>[1 3 7] |

Tabla 4.9: Resultados de la multiclasificación con Random Forest del conjunto de datos MIAS

El valor resultante para F1-score fue de 62 % manteniendo un valor de 62 % para la precisión y un valor de 63 % en exactitud y recall. También, su matriz de confusión se encuentra mayormente ponderada en los casos correctamente clasificados.

SVM es el clasificador de menor rendimiento en comparativa con los aquí presentados, situación que se puede visualizar en la Tabla 4.10.

| SVM | | | | | |
|---|---|---|---|---|---|
| **Fold** | **F1-score** | **Precision** | **Accuracy** | **Recall** | **Confusion matrix** |
| 1 | 0.14215686274509807 | 0.15187165775401067 | 0.18235294117647056 | 0.18235294117647056 | [ 0 0 11]<br>[ 0 1 10]<br>[ 0 0 12] |
| 2 | 0.15414322250639384 | 0.14456747404844291 | 0.15294117647058826 | 0.15294117647058826 | [ 0 0 11]<br>[ 0 0 11]<br>[ 0 0 12] |
| 3 | 0.10040125885129816 | 0.13010752688172042 | 0.1548387096774194 | 0.13548387096774194 | [0 1 9]<br>[ 0 0 10]<br>[ 0 0 11] |
| 4 | 0.16566666666666666 | 0.1111111111111111 | 0.3333333333333333 | 0.3333333333333333 | [ 0 0 11]<br>[ 0 0 11]<br>[ 0 0 11] |
| 5 | 0.15587209302325582 | 0.1481640625 | 0.14375 | 0.14375 | [ 0 0 11]<br>[ 0 0 10]<br>[ 0 0 11] |
| 6 | 0.13687770562770563 | 0.1344758064516129 | 0.175 | 0.175 | [ 0 0 11]<br>[0 1 9]<br>[ 0 0 11] |
| **7** | **0.16666666666666666** | **0.1111111111111111** | **0.3333333333333333** | **0.3333333333333333** | [ 0 0 11]<br>[ 0 0 11]<br>[ 0 0 11] |
| 8 | 0.1558678955453149 | 0.1259105098855359 | 0.1548387096774194 | 0.1548387096774194 | [ 0 0 10]<br>[ 0 0 10]<br>[ 0 0 11] |
| 9 | 0.1458678955453149 | 0.1159105098855359 | 0.1448387096774194 | 0.1448387096774194 | [ 0 0 10]<br>[ 0 0 10]<br>[ 0 0 11] |
| 10 | 0.14040125885129816 | 0.13010752688172042 | 0.1548387096774194 | 0.1548387096774194 | [ 0 0 10]<br>[1 0 9]<br>[ 0 0 11] |

Tabla 4.10: Resultados de la multiclasificación con SVM del conjunto de datos MIAS

SVM arrojó un valor máximo de F1-score del 16 %, precisión del 11 %, exactitud y recall del 33 %. Así mismo, los elementos incorrectamente clasificados expresados en la matriz de confusión son mayores a aquellos correctamente clasificados. Con base a esto, SVM se coloca como el peor clasificador para el problema de multiclasificación de los niveles de densidad mamográfica para las imágenes MIAS.

Para el clasificador Naive Bayes los resultados se pueden visualizar en la Tabla 4.11.

| Naive Bayes | | | | | |
|---|---|---|---|---|---|
| **Fold** | **F1-score** | **Precision** | **Accuracy** | **Recall** | **Confusion matrix** |
| **1** | **0.6068134763786938** | **0.6090909090909091** | **0.6060606060606061** | **0.6060606060606061** | [8 0 3]<br>[2 6 3]<br>[1 4 6] |
| 2 | 0.6004201680672269 | 0.6062464985994398 | 0.6176470588235294 | 0.6176470588235294 | [8 1 2]<br>[1 8 2]<br>[2 5 5] |
| 3 | 0.5796638655462185 | 0.584841628959276 | 0.5764705882352942 | 0.5764705882352942 | [9 2 0]<br>[0 6 5]<br>[1 3 8] |
| 4 | 0.5909090909090908 | 0.5909090909090908 | 0.5909090909090908 | 0.5909090909090908 | [7 3 1]<br>[2 6 3]<br>[2 2 7] |
| 5 | 0.5219551282051281 | 0.5296875 | 0.53125 | 0.53125 | [6 3 2]<br>[2 3 5]<br>[1 2 8] |
| 6 | 0.5551136363636364 | 0.5690972222222222 | 0.55 | 0.55 | [8 3 0]<br>[1 7 2]<br>[0 2 9] |
| 7 | 0.580087315061848 | 0.5849136526555881 | 0.5774193548387096 | 0.5774193548387096 | [7 2 1]<br>[2 6 2]<br>[0 3 8] |
| 8 | 0.5555694987820181 | 0.5851736972704715 | 0.5451612903225806 | 0.5451612903225806 | [7 3 0]<br>[3 6 1]<br>[0 4 7] |
| 9 | 0.5841590780470237 | 0.596415770609319 | 0.5774193548387096 | 0.5774193548387096 | [8 2 0]<br>[1 6 3]<br>[0 4 7] |
| 10 | 0.6028054005982698 | 0.5982730531117627 | 0.5096774193548387 | 0.5096774193548387 | [10 0 0]<br>[1 5 4]<br>[0 4 7] |

Tabla 4.11: Resultados de la multiclasificación con Naive Bayes del conjunto de datos MIAS

Naive Bayes, expone un máximo de 60 % en F1-score, precisión, exactitud y recall. De igual forma, en la matriz de confusión aunque hay mayor número de casos correctamente clasificados no existe una variación drástica de resultados en los folds.

Los resultados implementando la base de datos MIAS con el programa genético se muestran en la Tabla 4.12.

| Our_GP | | | | | |
|---|---|---|---|---|---|
| **Fold** | **F1-score** | **Precision** | **Accuracy** | **Recall** | **Confusion matrix** |
| 1 | 0.6500754147812972 | 0.7245852187028659 | 0.6470588235294118 | 0.6470588235294118 | [8 3 0]<br>[3 8 0]<br>[2 4 6] |
| 2 | 0.4602683178534572 | 0.4617647058823529 | 0.47058823529411764 | 0.47058823529411764 | [7 3 1]<br>[1 2 8]<br>[2 3 7] |
| **3** | **0.7065656565656565** | **0.749417249417249** | **0.696969696969697** | **0.696969696969697** | [9 2 0]<br>[0 7 4]<br>[0 4 7] |
| 4 | 0.5810185185185185 | 0.7142857142857143 | 0.5757575757575758 | 0.5757575757575758 | [5 6 0]<br>[0 9 2]<br>[0 6 5] |
| 5 | 0.5285283521303258 | 0.5302350427350427 | 0.53125 | 0.53125 | [6 3 2]<br>[2 4 4]<br>[2 2 7] |
| 6 | 0.421335005574136 | 0.434375 | 0.4375 | 0.4375 | [6 2 3]<br>[2 2 6]<br>[4 1 6] |
| 7 | 0.6516129032258066 | 0.6792114695340502 | 0.6451612903225806 | 0.6451612903225806 | [6 4 0]<br>[1 7 2]<br>[1 3 7] |
| 8 | 0.5292220113851992 | 0.5678443420378904 | 0.5161290322580645 | 0.5161290322580645 | [4 6 0]<br>[3 5 2]<br>[0 4 7] |
| 9 | 0.6765594337750535 | 0.6913489736070381 | 0.6774193548387096 | 0.6774193548387096 | [9 1 0]<br>[2 6 2]<br>[0 5 6] |
| 10 | 0.5269268997921749 | 0.5585253456221199 | 0.5161290322580645 | 0.5161290322580645 | [5 5 0]<br>[1 5 4]<br>[1 4 6] |

Tabla 4.12: Resultados de la multiclasificación con el programa genético del conjunto de datos MIAS

El mejor fold de Our_GP expone un valor del 70 % con la métrica de F1-score, un 74 % en precisión, un 69 % en exactitud y recall. También, su matriz de confusión expresa que hay mayor número de casos correctamente clasificados en contraste con los incorrectamente clasificados.

Por otra parte, en la Figura 4.2, se puede ver el desempeño de Our_GP con base a la información especificada desde la Tabla 4.7, hasta la Tabla 4.12.

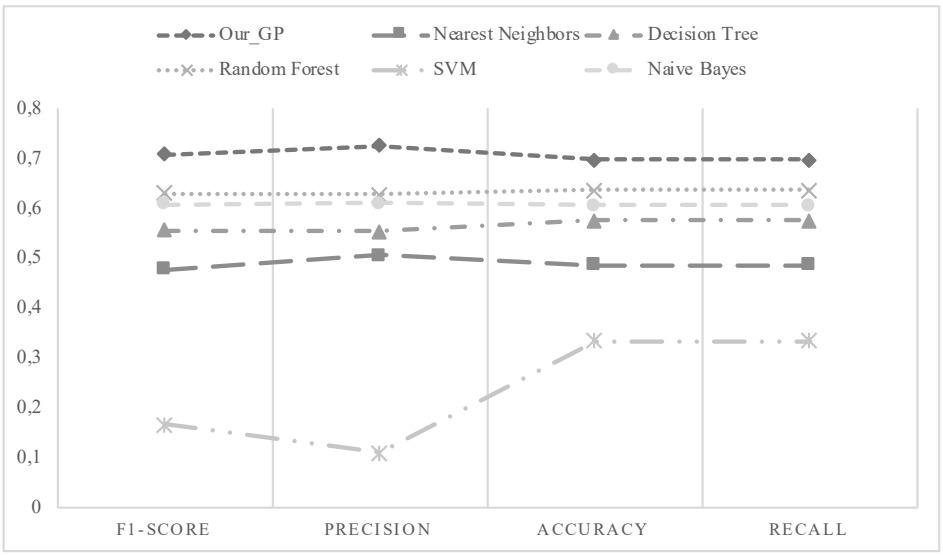


Figura 4.2: Gráfica de resultados de la multiclasificación con MIAS para todos los clasificadores

En la Figura 4.2 se puede identificar que Our_GP representa el mejor resultado, motivo por el que es posible decir que es competente para la solución del problema de multiclasificación de los niveles de densidad mamográfica.

## 4.3. Resultados del esquema de paralelización por islas

Los resultados de la paralelización de los niveles de densidad mamográfica se expone en las Tablas 4.13 y 4.14, donde se presenta el tiempo de ejecución y el valor del fitness (mediante F1-score) de cada una de las islas para ambos conjuntos de datos, INbreast y MIAS.

| No. Islas | Tiempo de ejecución | F1-score |
|---|---|---|
| **1** | 108.669s | 0.31544771018455225 |
| **2** | 82.2s | 0.3186429276654841 |
| **3** | 128.4s | 0.3061473127262601 |
| **4** | 74.4s | 0.29831583268516115 |
| **5** | 249s | 0.33524544118917987 |
| **6** | 128.4s | 0.2887092731829574 |
| **7** | 78.6s | 0.22697665573146963 |
| **8** | 83.4s | 0.33657407407407414 |
| **9** | 127.8s | 0.20590520590520592 |
| **10** | 128.4s | 0.2545051433136922 |

Tabla 4.13: Tiempo de ejecución y resultado de fitness por isla para INbreast

Tomando como referencia que el método de multiclasificación para las imágenes INbreast tarda 1272.6s en su ejecución normal, es posible identificar en la Tabla 4.13, que el método paralelizado mantiene un buen rendimiento durante la ejecución de cada una de las islas, con un promedio de 118.9s por isla.

| No. Islas | Tiempo de ejecución | F1-score |
|---|---|---|
| **1** | 203.4s | 0.6500754147812972 |
| **2** | 147s | 0.4602683178534572 |
| **3** | 197.4s | 0.7065656565656565 |
| **4** | 151.8s | 0.5810185185185185 |
| **5** | 144.6s | 0.5285283521303258 |
| **6** | 150s | 0.421335005574136 |
| **7** | 141s | 0.6516129032258066 |
| **8** | 196.2s | 0.5292220113851992 |
| **9** | 190.8s | 0.6765594337750535 |
| **10** | 88.8s | 0.5269268997921749 |

Tabla 4.14: Tiempo de ejecución y resultado de fitness por isla para MIAS

Ahora bien, para las imágenes MIAS el método de multiclasificación tarda 1807.8s en su ejecución secuencial, mientras que, en el ambiente paralelo cada una de las islas tarda

un promedio de 153.1s en ejecutarse, lo anteriormente mencionado se puede visualizar en la Tabla 4.14.

En la Tabla 4.15, se expone el método de multiclasificación para las imágenes de las bases de datos MIAS e INbreast, así como el tiempo de ejecución para 1, 8 y 16 workers, así como el Speedup de 1 a 8 y de 1 a 16 workers.

| **Método** | **1 worker** | **8 workers** | **16 workers** | **Speedup comparado con 8 workers** | **Speedup comparado con 16 workers** |
|---|---|---|---|---|---|
| **Multiclasificación de los niveles de densidad mamográfica con Programación Genética (INbreast).** | 1272.6s | 798s | 371.4s | 1.59s | 3.42s |
| **Multiclasificación de los niveles de densidad mamográfica con Programación Genética (MIAS).** | 1867.8s | 1345.2s | 1206.6s | 1.38s | 1.54s |

Tabla 4.15: Speedup del método en paralelo de la multiclasificación de los niveles de densidad mamográfica con Programación Genética

Para concluir, se obtuvo el rendimiento total del método de multiclasificación de los niveles de densidad mamográfica con ambas bases de datos (MIAS e INbreast), donde de forma secuencial se obtuvo un total de 1272.6s para INbreast, mientras que para MIAS fue 1867.8s. En cuanto al método en un ambiente paralelo se obtuvo 371.4s en INbreast y 1206.6s en MIAS. Por otra parte, el speedup comparado con 8 workers alcanzó un valor de 1.59s para INbreast y 1.38s para MIAS. Mientras que, el speedup comparado con 16 workers en INbreast fue de 3.42s y 1.54s en MIAS.

## CAPÍTULO 5

# Conclusiones y trabajo futuro

A lo largo de esta tesis fue posible identificar la magnitud del problema, comprendiendo que el cáncer de mama es una afección creciente para las mujeres, ya que representa la segunda causa de muerte en mujeres con cáncer, debido a la inoportuna detección y por esta razón es que se desarrolló un método de multiclasificación para el apoyo a la detección.

Principalmente, se logró hacer la extracción de características de las imágenes mamográficas correspondiente a cada una de las bases de datos analizadas, como lo son MIAS e INbreast, tal como se muestra en el Capítulo 3. También, durante el proceso de extracción una vez obtenido el ROI, se implementaron diversos métodos como: LBP, métodos estadísticos, análisis multiresolución y Haralick, donde una combinación de ellos nos brindó un mejor desempeño.

La principal propuesta fue crear un método de multiclasificación de los niveles de densidad mamográfica con Programación Genética, de tal manera que se logró obtener un resultado equiparable con los que ofrecen otros clasificadores, teniendo para imágenes INbreast una exactitud de 47 %, colocando al programa genético como el segundo mejor para la solución del problema de multiclasificación, solo después de Naive Bayes. Por otra parte, para las imágenes MIAS con el programa genético (Our_GP), se obtuvo un 70 % bajo la medida F1-score, haciéndolo el mejor multiclasificador para los niveles de densidad mamográfica.

Respecto a la paralelización, se logró reducir el tiempo de ejecución un 71 % para las imágenes de la base de datos INbreast, mientras que para las imágenes de la base de datos MIAS, el tiempo se redujo un 36 %. Así mismo, el desempeño del método de multiclasificación en el ambiente paralelo, arrojó para cada una de las islas una disminución considerable de tiempo, concluyendo que fue posible tener un mejor tiempo de ejecución para la multiclasificación de los niveles de densidades mamográficas con INbreast que con MIAS.

Finalmente, las preguntas de investigación quedan resueltas al demostrar que, primero, resultó altamente efectivo el pre-procesamiento de la imagen para la correcta extracción de las características, debido a que, al tener un ROI definido el margen de extracción queda

limitado al área de importancia descartando obtener valores que no aporten características competentes. Segundo, el método de multiclasificación de los niveles de densidad mamográfica a través de la Programación Genética si representa una mejora en exactitud respecto a las demás técnicas de clasificación, debido a que para INbreast se colocó como el segundo mejor, mientras que para MIAS fue la mejor solución. Y tercero, se efectuó una mejora en términos de tiempo y consumo de recursos con el método paralelizado, que disminuyó 901.2s el tiempo de ejecución en comparación con la aplicación secuencial.

Como parte del trabajo futuro, se pretende mejorar el rendimiento del método de multiclasificación de los niveles de densidad mamográfica con Programación Genética. Lo anterior partiendo de una mejora en el pre-procesamiento de la imagen y la aplicación de los métodos de extracción de características. Además, se pretende mejorar la selección de rangos para la multiclasificación dentro del programa genético.

# Productos obtenidos

## Publicación de un artículo en international journal of scientific and technical research in engineering (IJSTRE)

Durante el desarrollo del tema de tesis se hizo una publicación que se centra en la multiclasificación de los niveles de densidad mamográfica con programación genética, dicha publicación se podrá consultar en las memorias de IJSTRE mayo-junio de 2019, en el volúmen 4, con índice de publicaciones número 3. En el link http://ijstre.com/Publish/4302019/246886609.pdf. El aspecto del artículo publicado bajo el identificador número 246886609, se puede visualizar tal como se muestra en la Figura 5.

*International Journal of Scientific and Technical Research in Engineering (IJSTRE)*
*www.ijstre.com Volume 4 Issue 3 ǁ May-June 2019.*

# Multiclassification of mammographic density levels with genetic programming

Ninderlisbhet Vargas Cardenas[1], María Guadalupe Sánchez Cervantes[2], Daniel Fajardo-Delgado[3]
[1,2,3](Department of Systems and Computing, Instituto Tecnológico de Ciudad Guzmán, México)

***ABSTRACT:*** *Breast cancer is the most frequent in women, as well as cervical and skin cancer. Medical research for the prevention of breast cancer has shown that breast density is a strong indicator of cancer risk. The density can be evaluated through the classification proposed by the American College of Radiology (ACR). The objective of this article is to show the results obtained from the multiclassification of mammographic density levels using genetic programming. The multiclass classification starts from a set of texture characteristics of the mammographic images. Various methods of feature extraction and classification have been implemented. The features of the images are the input for the classifiers. For the experiments, the INbreast mammography database was used. The results show good classification and will help doctors to make a reliable diagnosis of cancer risk.*

*{vargascardenasn, magusace, dfajardod}@gmail.com*

***KEYWORDS***–*Breast Density, Genetic Programming, Multiclassification.*

## I. INTRODUCTION

Breast cancer is a disease with a complex natural evolution that, despite the advances of modern oncology, is the second cause of death by neoplasia in women worldwide, with nearly 500 thousand deaths each year, from which 70% occurs in developing countries [1]. Breast cancer can be detected through microcalcifications and mammographic breast densities. Approximately 55% of non-palpable breast cancers present with visible microcalcifications. For better study over time, some methods of mammary density classification have been implemented, such as the "TNM" system of the American Joint Committee on Cancer (AJCC), and the BIRADS classification system (Breast Imaging Reporting and Data Systems). However, the detection of breast cancer with conventional methods applied by specialists has not been enough, now some dynamic methods have been implemented. In the process of providing a solution, emphasis has been placed on various classification techniques and, based on this, [2] they investigated the evolutionary algorithms of Genetic Programming (GP) and Learning Classification Systems (LCS), that work in such a way that they allow to classify the mammographic images by their levels of density under a binary classification environment. However, the proposed method allows a multiclass classification of mammographic density levels. To carry out this classification through GP, there are a series of steps to follow, such as the pre-processing of the image, the extraction of the characteristics and the implementation of GP for the multiclassification.

## II. ENVIRONMENT

This research is based on various concepts such as the ones mentioned below, which are important for the understanding of this article.

**2.1. Binary classification and multiclassification**



Figura 5.1: International Journal of Scientific and Technical Research in Engineering. Volume 4 Issue 3, May-June 2019.

# Glosario

La esencia del proyecto está basada en diversos conceptos como los que se enlistan a continuación de manera abstracta, pero que son importantes para la realización del mismo:

- Cáncer de mama:

  El cáncer de mama es una enfermedad con una evolución natural compleja por lo que, a pesar de los avances de la oncología moderna, es la primera causa de muerte por neoplasia en la mujer en el ámbito mundial,con cerca de 500 mil muertes cada año, de las cuales el 70 % ocurre enpaíses en desarrollo (Cárdenas et al., 2013).

- Detección de cancer de mama:

  La detección actualmente se genera con el diagnóstico por imagen, que permite visualizar anomalías por diversas técnicas como la mastografía, el ultrasonido mamario, una resonancia magnética o por tomografía por emision de positrones (J. Cárdenas Sánchez et al., 2013).

- Imágenes Mamográficas: Las técnicas de diagnóstico por imagen se comprende por:

  1. Mastografía. Es el único método de imagen que ha demostrado disminución en la mortalidad por cáncer de mama al permitir un diagnóstico temprano.
  2. Ultrasonido (US) mamario. Valiosa herramienta complementaria de la mastografía diagnóstica, no útil como método de tamizaje para cáncer. Se requieren equipos de alta resolución, así como experiencia y conocimiento de la anatomía de la glándula mamaria y su evaluación por ecografía.
  3. Resonancia magnética (RM). Otro método de imagen que no utiliza radiación ionizante y proporciona información morfológica y funcional, a través de la inyección endovenosa de una sustancia paramagnética (gadolinio).
  4. Tomografía por emisión de positrones (PET CT). Es un estudio que combina tomografía computada (CT) con medicina nuclear (PET) en una misma imagen y permite en forma simultánea un estudio no sólo morfológico sino también funcional (metabólico) para la localización exacta de metástasis (J. Cárdenas Sánchez et al., 2013).

- Métodos de clasificación:

  El método de clasificación más común para clasificar el cáncer de mama es el sistema "TNM" del AJCC (American Joint Committee on Cancer).

  - T se refiere al tamaño del tumor. El valor de T es de entre 0 y 4, y describe el tamaño del tumor y la extensión a la piel o la pared torácica. Los valores de T más elevados indican la presencia de un tumor mayor o una mayor extensión a los tejidos cercanos a la mama.
  - N se refiere a los ganglios linfáticos. Los valores para N se sitúan entre 0 y 3, e indican si el cáncer se ha extendido a los ganglios linfáticos cercanos a la mama y, en caso de haberlo hecho, cuántos se encuentran afectados.
  - M se refiere a la metástasis. Los valores de M son 0 (ausencia de metástasis) ó 1 (presencia de metástasis), e indican si el cáncer se ha extendido a otros órganos distantes como los pulmones o los huesos.

  J. Cárdenas Sánchez et al. (2013) Mencionan que existe una variante denominada clasificación de BIRADS (Breast Imaging Reporting and Data Systems, American College of Radiology. Mammography, 4th ed., 2003).

- Niveles de densidad mamográfica:

  La densidad mamaria, estudiada a través de la mamografía (densidad mamográfica), refleja la composición del tejido mamario. El epitelio y estroma mamario producen mayor atenuación de los rayos X que la grasa, por lo que aparecen blancos en la mamografía, mientras que la grasa se ve oscura. Así la apariencia de la mamografía varía entre las mujeres, dependiendo de la composición de su mama. La proporción de mama constituida por tejido conectivo y epitelial es usualmente denominada como porcentaje de tejido mamario o Porcentaje de Densidad Mamográfica (PDM) (P.Neira, 2012).

- Sistemas de clasificación de aprendizaje:

  Los LCS evolucionan un conjunto de reglas incrementalemnte, a partir de un muestreo uniforme de los ejemplos del conjunto d enetrenamiento. Se ha demostrado que la complejidad del aprendizaje de cada regla depende del número de ejemplos representativos de la misma. Es decir, el sistema aprende de forma más rápida de las regiones de clasificación que cuenta con más ejemplos representativos (E. Bernadó Mansilla, 2004).

- Programación genética:

  La Programación Genética, PG, es un retoño de los Algoritmos Genéticos, en la cual los cromosomas que sufren la adaptación son en sí mismos programas de computador.

Se usan operadores genéticos especializados que generalizan la recombinación sexual y la mutación, para los programas de computador estructurados en árbol que están bajo adaptación. La PG trata de resolver uno de las cuestiones más excitantes e interesantes de las ciencias de la computación: ¿cómo pueden aprender los computadores a solucionar problemas sin que se les programe explícitamente? En otras palabras, la cuestión es cómo se puede hacer para que los computadores hagan lo que tienen que hacer, sin necesidad de la intervención humana que les diga exactamente como lo deben hacer(J. J. Martínez Páez, 2001).

- Cómputo paralelo:

  El cómputo paralelo es una forma de procesamiento que permite dividir una tarea en varias partes, de esta forma cada una de ellas se procesa de forma simultánea, utilizando hardware (Hw) con capacidades suficientes, más no potentes, mejorando el tiempo de respuesta del procesamiento (H. G. Sierra Varela, 2011).

- Cómputo secuencial

  Turing fue el visionario que propuso el modelo de cómputo secuencial mediante el cual una máquina compuesta de una cabeza lectora, una cinta infinitamente larga y un conjunto finito de estados, podía resolver un problema lógico, de manera algorítmica a lo que se define como el nacimiento del computador secuencial, mismo que llamó Máquina de Estado Finito (MEF) que de forma evidente implementa el cómputo secuencial por la forma de resolver el problema lógico (A. Turing, 1950).

- Cómputo Heterogéneo

  El cómputo heterogéneo hace referencia a la creación de sistemas que tienen la capacidad de usar más de un tipo de procesador o arquitecturas con la finalidad de mejorar el rendimiento, por lo general se incorporan capacidades de procesamiento especializadas para tareas en particular (A. Shan, 2006).

- Taxonomía de Flynn

  Con el paso de los años se han propuesto diversos esquemas de clasificación para los sistemas de cómputo con varios CPU, pero ninguno de ellos ha tenido un éxito completo ni se ha adoptado de manera amplia. Es probable que la taxonomía más citada sea la de Flynn (1972) aunque es algo rudimentaria Flynn eligió dos características consideradas por él como esenciales el número de flujos de instrucciones y el número de flujos de datos. Una computadora con un flujo de instrucciones y uno de datos se llama SISD (Single Instruction Single Data) Todas las computadoras tradicionales de un procesador (es decir aquellas que tienen un CPU) caen dentro de esta categoría desde las

computadoras personales hasta las grandes mainframes. La siguiente categoría es SIMD (Single Instruction Multiple Data) con un flujo de instrucciones y varios flujos de datos. Este tipo se refiere a ordenar procesadores con unidad de instrucción que busca una instrucción y después instruye a varias unidades de datos para que la lleven a cabo en paralelo cada una con sus propios datos. Estas máquinas son útiles para los cómputos que repiten los mismos cálculos en varios conjuntos de datos. Por otra parte, está la categoría MISD (Multiple Instruction Single Data) con un flujo de varias instrucciones y un flujo de datos. Ninguna de las computadoras conocidas se ajusta a este modelo y por último está MIMD (Multiple Instruction Multiple Data) que significa un grupo de computadoras independientes cada una con su propio contador del programa y datos. Todos los sistemas distribuidos son MIMD (A. S. Tanenbaum, 1996).

- Programación Multi-GPU

  Un sistema multi-GPU es aquel compuesto por varias GPUs que permiten amortizar el consumo de energía de un nodo del servidor a través de la GPU al ofrecer más rendimiento para una unidad dada de energía consumida, que al mismo tiempo aumenta su rendimiento. Es así que, al convertir una aplicación para aprovechar múltiples GPUs, es importante diseñar correctamente la comunicación inter-GPU. La eficiencia de los transceptores de datos inter-GPU depende de cómo las GPUs están conectadas dentro de un nodo, y en un clúster (J. Cheng y M. Grossman, T. McKercher, 2014).

- Programación Dinámica

  La programación dinámica (DP) es una técnica comúnmente utilizada para resolver una amplia variedad de problemas discretos de optimización, como programación, edición de cadenas, empaquetado y administración de inventario. DP ve un problema como un conjunto de subproblemas interdependientes. Resuelve subproblemas y usa los resultados para resolver subproblemas más grandes hasta que se resuelve todo el problema. A diferencia de Dividir-y-vencer, donde la solución a un problema depende solo de la solución a sus subproblemas, en DP puede haber interrelaciones entre los subproblemas. En DP, la solución a un subproblema se expresa como una función de soluciones a uno o más subproblemas en los niveles precedentes (V. Kumar et al., 1994).

- Costo computacional

  Existen problemas capaces de resolver en tiempo polinomial, es decir, para entradas de tamaño n, su peor tiempo de ejecución es $O(n^k)$ para una constante k. Pero, no todos los problemas se pueden resolver en tiempo polinomial, sin embargo, es una clase interesante de problemas, llamados problemas "NP-completos", cuyo estado es

desconocido. Aún no se ha descubierto ningún algoritmo de tiempo polinomial para un problema NP-completo, ni nadie ha podido probar que no pueda existir un algoritmo de tiempo polinomial para ninguno de ellos (T. H. Cormen et al., 2009).

- Velocidad de Procesamiento

  El procesador está formado por un conjunto de registros que almacenen datos, una unidad aritmético-lógica que realiza operaciones con ellos y una unidad de control que se encarga de coordinar a todos los componentes. Debido a que las operaciones dentro del procesador se sincronizan con los tics del reloj, la velocidad máxima del ordenador vendrá marcada por el ritmo de oscilación del reloj interno (G. Beekman, 2005).

# Referencias Bibliográficas


Alonso, J. M. (1997). Programación de aplicaciones paralelas con MPI (Message Passing Interface).

Aly, M. (2005). Survey on multiclass classification methods. Survey on Multiclass Classification Methods, Technical Report, Caltech, USA, 2005.

Arce, S. (2018). Implementación de un programa genético para niveles de densidad mamográfica. Instituto tecnológico de Cd. Guzmán, México.

Shan, A. (2006). Heterogeneous processing: a strategy for augmenting moore's law. Linux Journal, 2006(142), 7.

American College of Radiology. (1992). Breast Imaging Reporting and Data System® (BI-RADS®). American College of Radiology, Reston, Va.

Arancibia, P., Taub, T., y Grazia, K. De. (2013). Microcalcificaciones mamarias: revisión de los descriptores y categorías BI-RADS. Rev. Chil. Obstet. Ginecol., 78(5), 383–394.

Beekman George.(2005). Introducción A La Informática. Sexta edición Pearson Educación, S.A., Madrid.

Bernadó-Mansilla, E. (2004). Complejidad del aprendizaje y muestreo de ejemplos en sistemas clasificadores. In Proceedings del III Congreso Español de Metaheurísticas, Algoritmos Evolutivos y Bioinspirados (pp. 203-210).

Blot, L., y Zwiggelaar, R. (2001, July). Background texture extraction for the classification of mammographic parenchymal patterns. In Medical Image Understanding and Analysis (pp. 145-148).

Bosch, A., Munoz, X., Oliver, A., y Marti, J. (2006, June). Modeling and classifying breast tissue density in mammograms. In 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06) (Vol. 2, pp. 1552-1558). IEEE.

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

Burling-Claridge, F., Iqbal, M., y Zhang, M. (2016). Evolutionary Algorithms for Classification of Mammographic Densities using Local Binary Patterns and Statistical Features. Proceedings of 2016 IEEE Congress on Evolutionary Computation (CEC 2016), 3847–3854. doi: https://doi.org/doi:10.1109/CEC.2016.7744277

Burt, P. J., y Adelson, E. H. (1983). A multiresolution spline with application to image mosaics. ACM transactions on Graphics, 2(4), 217-236.

Cárdenas Sánchez, J. (2013). Consenso Nacional sobre tratamiento del cáncer mamario. Rev Inst Nac Cancerol (Mex), 41(3), 136-43.

Cheng, J., Grossman, M., y McKercher, T. (2014). Professional Cuda C Programming. John Wiley and Sons.

Chow, E., Falgout, R. D., Hu, J. J., Tuminaro, R. S., y Yang, U. M. (2006). A survey of parallelization techniques for multigrid solvers. In Parallel processing for scientific computing (pp. 179-201). Society for Industrial and Applied Mathematics.

Cormen, T. H., Leiserson, C. E., Rivest, R. L., y Stein, C. (2009). Introduction to algorithms. MIT press.

Cristianini, N., y Shawe-Taylor, J. (2000). An introduction to support vector machines and other kernel-based learning methods. Cambridge university press.

Daubechies, I. (1992). Ten lectures on wavelets (Vol. 61). Siam.

Espejo, P. G., Ventura, S., y Herrera, F. (2010). A Survey on the Application of Genetic Programming to Classifcation. Ieee Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, 40(2), 121–144.

Flynn, M. J. (1972). Some computer organizations and their effectiveness. IEEE transactions on computers, 100(9), 948-960.

Frank Eibe, Hall Mark. (2001). A Simple Approach to Ordinal Classification. Lecture Notes in Computer Science. 2167. 145-156. 10.1007/3-540-44795-4_13.

Haralick, R. M., y Shanmugam, K. (1973). Textural features for image classification. IEEE Transactions on systems, man, and cybernetics, (6), 610-621.

Hold-Geoffroy, Y., Gagnon, O., y Parizeau, M. (2014, July). Once you SCOOP, no need to fork. In Proceedings of the 2014 Annual Conference on Extreme Science and Engineering Discovery Environment (p. 60). ACM.

Ingalalli, V., Silva, S., Castelli, M., y Vanneschi, L. (2014, April). A multi-dimensional genetic programming approach for multi-class classification problems. In European Conference on Genetic Programming (pp. 48-60). Springer, Berlin, Heidelberg.

Karsavuran, M. O., Akbudak, K., y Aykanat, C. (2016). Locality-Aware Parallel Sparse Matrix-Vector and Matrix-Transpose-Vector Multiplication on Many-Core Processors. IEEE Transactions on Parallel and Distributed Systems, 27(6), 1713–1726.

Keller, J. M., Gray, M. R., y Givens, J. A. (1985). A fuzzy k-nearest neighbor algorithm. IEEE transactions on systems, man, and cybernetics, (4), 580-585.

Kriti, y Virmani, J. (2015). Breast density classification using Laws mask texture features. International Journal of Biomedical Engineering and Technology, 19(3), 279-302.

Knaul, F. M., Nigenda, G., Lozano, R., Arreola-Ornelas, H., Langer, A., y Frenk, J. (2008). Breast cancer in Mexico: a pressing priority. Reproductive Health Matters, 16(32), 113–123. doi: https://doi.org/10.1016/S0968-8080(08)32414-8

Koza, J. R., y Koza, J. R. (1992). Genetic programming: on the programming of computers by means of natural selection (Vol. 1). MIT press.

Kumar, V., Grama, A., Gupta, A., y Karypis, G. (1994). Introduction to parallel computing: algorithm design and analysis.

Latter, B. D. H. (1973). The island model of population differentiation: a general solution. Genetics, 73(1), 147-157.

Li, J. B. (2012). Mammographic image based breast tissue classification with kernel self-optimized fisher discriminant for breast cancer diagnosis. Journal of medical systems, 36(4), 2235-2244.

Loveard, T., y Ciesielski, V. (2001). Representing classification problems in genetic programming. Proceedings of the 2001 Congress on Evolutionary Computation (IEEE Cat. No.01TH8546), 2, 1070–1077.

Markatou, M., Tian, H., Biswas, S., Hripcsak, G. (2005). Analysis of variance of cross-validation estimators of the generalization error. J. Mach. Learn. Res., 6:1127–1168 (electronic). MR2249851.

Melab, N., y Talbi, E. G. (2010, July). GPU-based island model for evolutionary algorithms. In Proceedings of the 12th annual conference on Genetic and evolutionary computation (pp. 1089-1096). ACM.

Muhimmah, I., y Zwiggelaar, R. (2006, October). Mammographic density classification using multiresolution histogram information. In Proceedings of the International Special Topic Conference on Information Technology in Biomedicine, ITAB (pp. 26-28).

Muni, D. P., Pal, N. R., y Das, J. (2004). A novel approach to design classifiers using genetic programming. IEEE transactions on evolutionary computation, 8(2), 183-196.

Mustra, M., Grgic, M., y Delac, K. (2012). Breast density classification using multiple feature selection. automatika, 53(4), 362-372.

Neira Paulina, V. (2012). Densidad mamaria y riesgo de cáncer mamario. Revista Médica Clínica Las Condes, 24(1), 122–130. https://doi.org/10.1016/S0716-8640(13)70137-8

NVIDIA. ¿Qué es CUDA?. CUDA y el GPU Computing. 2017. http://www.nvidia.es/object/cuda-parallel-computing-es.html.

Ojala, T., Pietikainen, M., y Harwood, D. (1996). A comparative study of texture measures with classification based on featured distributions. Pattern recognition, 29(1), 51-59.

OpenMP. OpenMP Application Programming Interface. Versión 4.5. OpenMP Architecture Review Board, 2015.

Otsu, N. (1979). A threshold selection method from gray-level histograms. IEEE transactions on systems, man, and cybernetics, 9(1), 62-66.

Páez, J. J. M. (2001). Conceptos básicos de programación genética. Revista de la Facultad de Medicina, 49(2), 110-114.

Reyad, Y. A., Berbar, M. A., y Hussain, M. (2014). Comparison of statistical, LBP, and multi-resolution analysis features for breast mass classification. Journal of medical systems, 38(9), 100.

Romero-Laorden, D., Villazon-Terrazas, J., Martinez-Graullera, O., Ibanez, A., Parrilla, M., y Penas, M. S. (2016). Analysis of Parallel Computing Strategies to Accelerate Ultrasound Imaging Processes. IEEE Transactions on Parallel and Distributed Systems, 27(12), 3429–3440.

Saftlas, A. F., Hoover, R. N., Brinton, L. A., Szklo, M., Olson, D. R., Salane, M., y Wolfe, J. N. (1991). Mammographic densities and risk of breast cancer. Cancer, 67(11), 2833–2838. doi: https://doi.org/10.1002/1097-0142(19910601)67:11¡2833

Sharma, V., y Singh, S. (2014). CFS-SMO based classification of breast density using multiple texture models. Medical and biological engineering and computing, 52(6), 521-529.

Sharma, V., y Singh, S. (2015). Automated classification of fatty and dense mammograms. Journal of Medical Imaging and Health Informatics, 5(3), 520-526.

Sheshadri, H. S., y Kandaswamy, A. (2007). Experimental investigation on breast tissue classification based on statistical feature extraction of mammograms. Computerized Medical Imaging and Graphics, 31(1), 46-48.

Sierra Varela Heidi Gabriela. (2011). Cómputo Paralelo. Editorial Académica Española. (1):6-10

Silva, W. R., y Menotti, D. (2012). Classification of mammograms by the breast composition. In Proceedings of the International Conference on Image Processing, Computer Vision, and Pattern Recognition (IPCV) (p. 1). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp).

Smart, W., y Zhang, M. (2003). Classification strategies for image classification in genetic programming. Proceeding of Image and Vision Computing Conference, 402–407.

Smart, W. R. (2005). Genetic Programming for Multiclass Object Classification. Proceeding of Image and Vision Computing Conference, 402–407.

Society, T. I., Engineering, O., Source, O., Health, E., Alliance, R., Source, O.,View, R. (2017). networks and genetic algorithms. doi: https://doi.org/ 10.1117/12.274136

Subashini, T. S., Ramalingam, V., y Palanivel, S. (2010). Automated assessment of breast tissue density in digital mammograms. Computer Vision and Image Understanding, 114(1), 33-43.

Swain, P. H., y Hauska, H. (1977). The decision tree classifier: Design and potential. IEEE Transactions on Geoscience Electronics, 15(3), 142-147.

Tanenbaum, A. S., Guerrero, G., y Velasco, Ó. A. P. (1996). Sistemas operativos distribuidos (No. QA76. 76063. T35. 3 1996.). México;: Prentice Hall.

Turing Alan. (1950). "Computing Machinery and Intelligence" En: Mind, 59.

Tzikopoulos, S. D., Mavroforakis, M. E., Georgiou, H. V., Dimitropoulos, N., y Theodoridis, S. (2011). A fully automated scheme for mammographic segmentation and classification based on breast density and asymmetry. computer methods and programs in biomedicine, 102(1), 47-63.

Virmani, J. (2015). Breast tissue density classification using wavelet-based texture descriptors. In Proceedings of the Second International Conference on Computer and Communication Technologies (pp. 539-546). Springer, New Delhi.

Virmani, J. (2016). Comparison of CAD systems for three class breast tissue density classification using mammographic images. In Medical imaging in clinical applications (pp. 107-130). Springer, Cham.

Virmani, J., Dey, N., y Kumar, V. (2016). PCA-PNN and PCA-SVM based CAD systems for breast density classification. In Applications of intelligent optimization in biology and medicine (pp. 159-180). Springer, Cham.

Virmani, J., y Thakur, S. (2016). Application of statistical texture features for breast tissue density classification. In Image Feature Detectors and Descriptors (pp. 411-435). Springer, Cham.

Wu, Y., Giger, M. L., Doi, K., Vyborny, C. J., Schmidt, R. A., & Metz, C. E. (1993). Artificial neural networks in mammography: application to decision making in the diagnosis of breast cancer. Radiology, 187(1), 81-87.

Whitley, D., Rana, S., y Heckendorn, R. B. (1998). The Island Model Genetic Algorithm: On Separability. Population Size and Convergence, 256.

Zemmal, N., Azizi, N., Sellami, M., y Dey, N. (2016, July). Breast abnormalities diagnosis based on transductive SVM a mutual information based feature selection. In Proceedings of international conference for engineering and sciences, Barcelona (pp. 23-27).

Zhang, M., y Smart, W. (2004, April). Multiclass object classification using genetic programming. In Workshops on Applications of Evolutionary Computation (pp. 369-378). Springer, Berlin, Heidelberg.

Zhang, H. (2004). The optimality of naive Bayes. AA, 1(2), 3.