Diseño e implementación de un método automático de clasificación de linfocitos afectados por leucemia linfoblástica aguda en imágenes hematológicas.

Sebastian Felipe Pinto González

Universidad El Bosque
Facultad de ingeniería, Programa de Bioingeniería
Bogota D.C., Colombia
2018

Diseño e implementación de un método automático de clasificación de linfocitos afectados por leucemia linfoblástica aguda en imágenes hematológicas.

Sebastian Felipe Pinto Gonzalez

Tesis o trabajo de grado presentada(o) como requisito parcial para optar al título de:
Bioingeníero

Director(a):
Jhonathan Tarquino González

Bogota D.C., Colombia
2018

Contenido

Lista de Figuras

Lista de Tablas

1 Resumen

En Colombia la leucemia linfoblástica aguda (LLA) es una enfermedad de alta prevalencia en menores de edad. Esta es en muchos casos mortal si no se diagnostica a tiempo y dada su alta variabilidad sintomatológíca es complejo que este sea temprano; uno de los exámenes que permiten su diagnóstico es realizar estudios hematológicos por medio de muestras de sangre, dentro de estos estudios se encuentra la inspección visual de la sangre, la cual hace uso de imágenes hematológicas para dar un resultado, el proceso de inspección visual al ser un proceso manual y repetitivo se encuentra afectado por la fatiga del profesional especializado provocando una subjetividad y variabilidad que puede llegar a afectar el diagnóstico entregado.

Este último proceso puede ser beneficiado con la inclusión de herramientas que permitan ayudar al profesional en su tarea del diagnóstico, por medio del análisis y procesamiento de imágenes, el cual puede extraer y resaltar características de las imágenes que usualmente son poco visibles para el ojo humano, y con estas características realizar estudios diferenciales entre las características de muestras de LLA positivas y LLA negativas.

En este trabajo de grado se hace un acercamiento al método manual realizado por los especialista por medio del análisis y procesamiento de imágenes, el cual permite extraer objetos y características específicas, permitiendo crear grupos de características para muestras LLA positivas y LLA negativas, las cuales pueden ser analizadas e introducidas a algoritmos de predicción; en este caso se usó una máquina de soporte vectorial, para tener un método completo para apoyar al especialista.

El método realizado se validó usando diferentes métricas como lo son la sensibilidad, especificidad, exactitud y precisión, las cuales dieron índices por encima del 70 %, estos valores son confirmados y soportados por una curva ROC la cual dio un valor superior al 75 %.

2 Introducción

La leucemia linfoblástica aguda (LLA) es una enfermedad de alta prevalencia en población Colombiana menor de 17 años, teniendo una incidencia de 764 casos en niños y 558 casos en niñas en el año 2015 [Jairo Aguilera López, 2015] , convirtiéndose de esta forma en un problema de salud pública [del pilar, 2016]. Para el diagnóstico de esta enfermedad se realizan diferentes pruebas médicas que incluyen inspección visual de muestras hematológicas por microscopía óptica, evaluación de muestras histológicas, pruebas de los cromosomas, aspiración medular y biopsia de la médula ósea, punción lumbar, biopsia de los ganglios linfáticos, citometría de flujo e inmunohistoquímica principalmente; todas estas con el objetivo de descartar otras enfermedades con sintomatología similar [Society, 2016].

Para el estudio de pacientes con posible LLA los oncólogos especializados deben examinar las muestras en centros de oncohematología para poder evaluar el estado de los órganos que originan la sangre y determinar el contenido de la misma mediante inspección por microscopía óptica [Gersten, 2018]. Este tipo de evaluación de muestras sanguíneas se realiza de manera manual, por lo que el ejercicio se ve afectado por la fatiga que provoca el procedimiento por ser una tarea repetitiva [Hamid, 2013, Putzu L, 2014] y por los diferentes niveles de entrenamiento del especialista. Es así que esta tarea de análisis de imágenes médicas se ve influenciada por factores que incrementan la subjetividad y variabilidad en el diagnóstico de este tipo de Leucemia [Kandil, 2016, H, 2012].

Además, debido a la mencionada subjetividad y variabilidad de conceptos médicos se ve limitada la efectividad de posibles tratamientos y se condiciona el pronóstico del paciente [Amin M, 2015]. Con este panorama, el desarrollo de herramientas automáticas que asistan el diagnóstico a partir de imágenes hematológicas aparece como un método para disminuir las posibles consecuencias de errores en conceptos médicos y también provee una aceleración del flujo de trabajo donde el análisis manual de imágenes se convierte en cuello de botella dentro de los procesos clínicos [Kandil, 2016].

2.1. Estado del arte

Partiendo del estado del arte en herramientas CAD (Diagnóstico Asistido por Computadora), este problema de desarrollo de técnicas computarizadas de apoyo al especialista basadas en imágenes hematológicas ya ha sido atacado de diferentes formas, teniendo como punto de partida la detección de leucocitos. Para esta labor se han usado proyecciones en espacios de color alternos al RGB (Red,Green, Blue), como por ejemplo el HSV (Hud,Saturation,Value) [H, 2012] que permite separar los leucocitos del resto de contenido de las muestras en términos de intensidad, debido al contraste de fase que se halla entre el citoplasma y el núcleo [Miralles, 2017].

Posteriormente, se usaron umbrales de tamaño y forma [R., 2013] [S.Ordaz, 2011] que buscan descartar elementos que sean muy distantes a las características de los correspondientes de un leucocito disminuyendo el costo computacional. Sin embargo, este método ha mostrado tener limitantes en imágenes saturadas o con mucho ruido, por lo que se hace necesario aplicar filtros digitales junto con operadores morfológicos [Putzu L, 2014] que en conjunto disminuyen la variabilidad de las imágenes y la relación señal-ruido de las mismas [Mishra, 2017].

Una vez se ha diferenciado el leucocito y procesado la imagen para eliminar falsos positivos, varios autores intentaron caracterizar los leucocitos (sanos y LLA) mediante descriptores morfológicos basados en la forma del núcleo [S.Ordaz, 2011], el contraste entre núcleo-citoplasma [Mishra, 2017], diámetro y área del leucocito [Scottii, 2005], características que han demostrado ser discriminatorias de las clases LLA y sano. Sin embargo, a pesar de que estas características han mostrado resultados con una precisión 93 % también tienen un elevado costo computacional.

Caracterizados los leucocitos la literatura evidencia el uso de clasificadores como método de diferenciación entre los linfocitos sanos y enfermos (es decir células normales y cancerígenas). Lo anterior haciendo uso de las características seleccionadas y mencionadas anteriormente. Esta tarea de determinar de manera automática si los candidatos son sanos o no, se ha atacado mediante diferentes técnicas de clasificación representadas en la tabla **2-1**.

Los clasificadores son óptimos y eficientes siempre y cuando se realice un correcto proceso de entrenamiento y validación, por ende se usa un referente que indique la efectividad de los mismos, y para ello se usan bases de datos que proporcionan la información y las variables dentro del campo de aplicación. Es por esto que como parte de este proyecto de grado se ha decidido trabajar con bases de datos de imágenes que contengan anotaciones o etiquetas que faciliten el procesos de obtención de información diagnóstica por parte del especialista.

El desarrollo de este tipo de herramientas de asistencia requiere de conocimientos propios de la biología, la medicina y los procesos inmersos en el análisis matemático de imágenes, lo cuál evidencia la necesidad de un profesional con formación en estas temáticas como gestor de una solución que tenga en cuenta aspectos de todas las áreas mencionadas. Dicho perfil se alinea con el del Bioingeniero de la Universidad El Bosque y de manera más explícita, la temática del proyecto está dentro del foco misional de tecnologías para la salud de entes biológicos pues desarrolla una metodología para apoyar el diagnóstico de padecimientos propios del ser humano.

2.2. Objetivos

Dado lo anterior, se plantea diseñar e implementar un método para la clasificación de linfocitos afectados por leucemia linfoblástica aguda en imágenes hematológicas como objetivo general y los siguientes como objetivos específicos:

- Diseñar un método para la identificación automática de linfocitos a partir de imágenes hematológicas.
- Determinar las características que discriminan linfocitos afectados por leucemia linfoblástica aguda en imágenes de microscopía óptica.
- Diseñar un algoritmo de clasificación automática linfocitos afectados por leucemia linfoblástica aguda en imágenes de microscopía óptica, adaptado para las características discriminantes encontradas.
- Implementar en un lenguaje de programación el método automático de clasificación diseñado.
- Validar el método implementado para clasificación de linfocitos afectados por leucemia linfoblástica aguda en imágenes de microscopía óptica.

La estructura de este documento se encuentra dividida de la siguiente forma: en el segundo capítulo se expondrá un marco teórico que explicará los elementos técnicos y teóricos usados en el documento para facilitar su entendimiento y se dan ejemplos de su uso y funcionamiento.

En el tercer capítulo se encuentra la metodología usada con muestras de la forma como se condujo el trabajo de grado paso a paso y se evidencian los resultados de la metodología planteada, en el cuarto capítulo se presentan las conclusiones y recomendaciones del proyecto.

Se aclara que el proyecto de grado planteado es una aproximación preliminar de herramienta para apoyo al especialista en diagnóstico de Leucemia Linfoblástica Aguda (LLA), por lo que para los pacientes debe resultar transparente el funcionamiento de la misma.

MÉTODO DE CLASIFICACIÓN	VENTAJAS	DESVENTAJAS
Sistemas inmunes artificiales	Sistema de optimizado automático, funciona por medio de afinidades y eliminación de los resultados menos afines, adaptable ante cambios en la muestra	Clasificación binaria y no supervisada, las muestras no afines son eliminadas y estas pueden ser falsos negativas sesgando el método [E.Cuevas, 2010, S.Ordaz, 2011]
Árboles de decisión	Manejo de datos discretos y continuos, permite un manejo de datos con costos diferenciales, crea sistemas de umbralización con base en los valores clasificados, puede funcionar con datos no numéricos	Funciona como un sistema lineal de cola para cada dato, en algunos casos se requiere de nodos no conectados para no adquirir ambigüedad en las muestras, aumentando la probabilidad de fallo. [Suca, 2016]
Maquinas de soporte vectorial	El uso de hiper-planos del sistema evita el error de sobre ajuste que tienen los demás métodos de clasificación, en lugar de minimizar la cantidad de errores en el entrenamiento del sistema, amplía el espacio entre los datos de manera casi infinita, es un método creado para la clasificación binaria que ha demostrado ser muy útil en otras aplicaciones.	Dependiendo del modelo del hiperplano usado altera la optimización del mismo, reduce los problemas no lineales en problemas lineales, la selección del kernel afecta de manera positiva o negativa el funcionamiento del sistema [Enrique J, 2014].
Redes neuronales	Tiene un sistema de aprendizaje amplio y de múltiples etapas, tiene tolerancia a cambios inesperados en el sistema en caso de que falle alguna parte del sistema, Es un método robusto que tolera fallos en los datos de entrada	El tiempo de aprendizaje es muy elevado y puede tener problemas de sobre ajuste, no tiene una salida nominal, por lo que los datos finales tienen que ser interpretados por un externo, se requiere de una cantidad extensa de datos para tener flexibilidad en la red neuronal, lo que incluye mas tiempo de entrenamiento. [Piuri, 2004, Mishra, 2017]

Tabla 2-1: Ventajas y desventajas de clasificadores usados como herramienta para el diagnóstico de LLA

3 Marco teórico

La sangre tiene como función principal transportar de oxígeno a nivel alveolar a través de la hemoglobina, además de funciones de coagulación y protección haciendo uso de los glóbulos blancos [Hamid, 2013]. Es un tejido líquido que provee de defensa al cuerpo ante bacterias, virus y cualquier agente microbiano o viral que sea considerado una amenaza. Para lograr su objetivo hace uso de células especializadas llamadas linfocitos, que evitan que los organismos amenazantes se dispersen por todo el cuerpo, atacándolos y eventualmente eliminándolos [Brummel, 2002].

Este fluido se presenta entre otros, cuatro compuestos principales a saber: glóbulos rojos (Eritrocitos), glóbulos blancos (Linfocitos), plasma y plaquetas [Brummel, 2002], cada uno de ellos con características distintivas.

- Los glóbulos rojos o eritrocitos, son células sin núcleo definido, son las células mas comunes y densas de la sangre, su principal función es la de transportar oxígeno al resto del cuerpo y retirar el CO2 (Dióxido de carbono) [Jiménez, 2008]. Los eritrocitos permiten realizar el intercambio gaseoso del cuerpo en el tejido alveolar por lo que son indispensables para mantener una salud estable.

- Los glóbulos blancos pueden ser clasificados en tres tipos a saber: linfocitos, monocitos y/o neutrófilos (Ver figura **3-1**), cada una de estas células con características y funciones diferentes. Específicamente, los neutrófilos son células especializadas para combatir bacterias, los monocitos proveen defensa instantánea engullendo y digiriendo todo tipo de amenaza al cuerpo y los linfocitos, que pueden ser linfocitos T o B dependiendo de su lugar de origen y de su modo de combatir las infecciones, producen anticuerpos altamente especializados como respuesta inmune o, secretan sustancias para atraer otras células inmunes para coordinar el ataque a la infección [Jiménez, 2008].

- Las plaquetas o trombocitos son células formadas en la médula ósea y tienen funciones principalmente coagulantes, actúan al momento de ruptura de cualquier vía de transporte sanguíneo.

- El plasma es el líquido mas abundante en la sangre y tiene la función de transportar nutrientes y desechos a otros organismos del cuerpo humano.

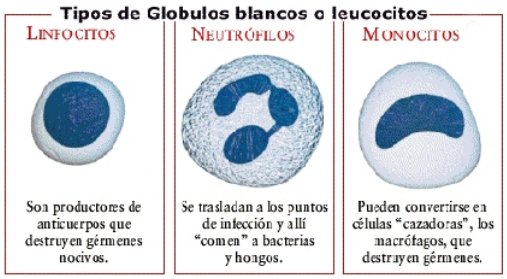

Figura 3-1: Tipos de glóbulos blancos [Lillo, 2012]

Como toda célula del cuerpo humano, los linfocitos tienen un ciclo de vida limitado, por lo que con el tiempo estos deben ser renovados, en los casos en que no son correctamente eliminados generan problemas al cuerpo produciendo linfocitos inmaduros carentes de utilidad. [Dr.Karthikeyan, 2017].

Además de lo mencionado anteriormente, los linfocitos pueden ser víctimas de un crecimiento excesivo e inmaduro que evita que funcionen correctamente, provocando que se acumulen y comiencen a atacar células no amenazantes, es decir, el cuerpo se ataca a sí mismo. Este comportamiento puede estar catalogado dentro de un marco hematopoyético de leucemia linfoblástica [Msalgobar, 2017, Dr.Karthikeyan, 2017, Jagadeesh, 2013]. Como se puede ver en la figura **3-2** las diferencias evidentes entre los linfocitos mostrados están en el tamaño del núcleo y del citoplasma, su morfología y la forma en la que responden a la tinción, aunque la forma de reacción a esta última no es un indicativo que sea positivo a leucemia. [Ariffin, 2012]

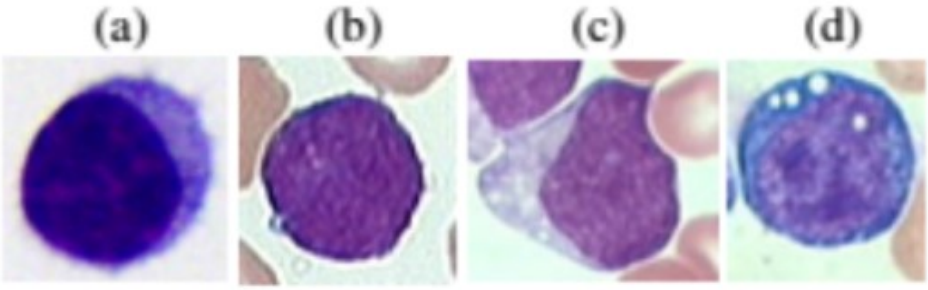

Figura 3-2: (a) Linfocito sano,(b) Linfocito infectado L1,(c) Linfocito infectado L2,(d) Linfocito infectado L3 [Piuri, 2004]

3.1. Tipos de leucemia

Existen dos tipos principales de leucemia, linfoblástica y mieloide, cada una de ellas con variantes (crónica y aguda), en referencia a la velocidad de su propagación [Brummel, 2002].

Es así que al presentarse una rápida dispersión en el cuerpo se le puede definir como leucemia aguda y por ende letal, pero cuando este crecimiento es lento o controlable se le llama crónica, que resulta ser tratable si se diagnostica a tiempo [Jiménez, 2008]. En la tabla **3-1** se puede apreciar las diferencias detalladas entre la leucemia linfoblástica aguda y la leucemia mieloide aguda.

AML(leucemia mieloide aguda)	ALL(leucemia linfoblastica aguda)
Afecta principalmente a adultos	Afecta a niños principalmente
Blastos pequeños en la célula	Blastos grandes en la célula
Exceso de citoplasma	Disminución de citoplasma
Se ven nucléolos entre 3-5 dentro del núcleo	nucléolos entre 1-3 dentro del núcleo
Presencia de gránulos en las células y varillas de auer	No hay gránulos
Tratamiento altamente tóxico	Baja toxicidad en el tratamiento
Alta tasa de mortalidad	Baja tasa de mortalidad
Se tiñe con myeloperoxidasa	Se usa tinción con ácido peróxido shiff

Tabla 3-1: Diferencias entre la leucemia mieloide aguda y la leucemia linfoblástica aguda [Hamid, 2013]

Para la clasificación de la leucemia linfoblástica aguda se usa el método francés-americabritánico (FAB) el cual clasifica los Linfoblástos en tres con las correspondientes características de cada uno, como se evidencia en la tabla **3-2**:

Esta clasificación al presentar irregularidades en el diagnóstico se reemplazó por otra que emplea técnicas moleculares donde también se evalúan precursores genéticos de las células y se reemplazó por LLA precursor B, LLA precursor T, ambos indicando cambios en el genoma de la célula [Sala, 2003]. Esta última clasificación al ser realizada por métodos moleculares no será usada, en su lugar se centrará el diagnóstico en el método de la FAB.

3.2. Diagnóstico de la leucemia linfoblastica

En el diagnóstico de la leucemia linfoblástica aguda se realizan varias pruebas clínicas incluyendo las nutricionales, psicológicas, físicas y de sangre, para descartar otros tipos de enfermedades que puedan tener síntomas similares a la leucemia debido a su amplio cuadro sintomático [Ofarrin, 2014]. En ocasiones se encuentra la leucemia durante la búsqueda de

	Categoría	Comentarios
LMA		
M0	LMA no diferenciada	Pobre Diagnóstico
M1	LMA con diferenciación mínima	Bastones de Aurer en blastos
M2	LMA con maduración	Pronóstico favorable en jóvenes
M3	Leucemia promielocítica	Blastos granulados; coaguloátias
M4	Leucemia mielomonocítica	Diferenciación mieloide/monocítica
M4co	Leucemia mielomonocítica eonsinofilia	Relativo buen pronóstico
M5a	Leucemia mielomonocítica, poco diferenciada	Enfermedad extramedular
M5b	Leucemia monocítica, bien diferenciada	Enfermedad extramedular
M6	Leucemia critroide	Rara; pronóstico pobre
M7	Leucemia megacariocítica	Rara; médula ósea fibrótica; pronóstico pobre
LLA		
L1	LLA infantis	Los blastos son células pequeñas con citoplasma pequeño
L2	LLA adulto	Los blastos son células grandes con citoplasma medio
L3	B-células maduras	Los blastos son células redondas con citoplasma basófilo

Tabla 3-2: Clasificación de leucemias agudas según Franco-americano-británico [Sala, 2003]

otras enfermedades, lo cual resulta común en niños, donde además de la dificultad asociada al diagnóstico, también se enfrentan al hecho de que no se puede examinar fácilmente todas las variables anteriormente nombradas sin un exhaustivo estudio del caso por parte del personal médico, por lo que se parte del estudio de sangre y ADN como complemento del estudio inicial , siempre buscando confirmación por medio de este último, en la figura **3-3** se nota la diferencia en el flujo sanguíneo [Amin M, 2015, Brummel, 2002, Kandil, 2016].

De los exámenes anteriormente mencionados, el estudio sanguíneo parte de un diagnóstico visual que usualmente se realiza de manera manual, teniendo en cuenta las características de forma, tamaño, relación entre citoplasma-núcleo, forma de núcleo y conteo total de linfocitos por campo visual. Este procedimiento es tedioso, lento y repetitivo, además presenta problemas al tener subjetividad y variabilidad inter e intra observador [H, 2017].

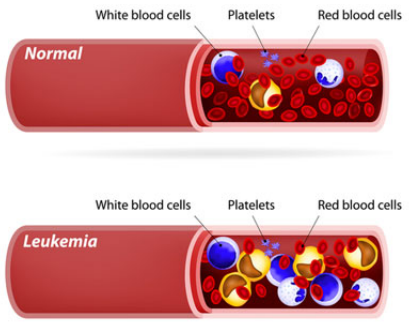

Figura 3-3: Presunción del efecto de la leucemia linfoblástica aguda en la sangre [PMfarma, 2015].

Todas las variables mencionadas en el párrafo anterior requieren de un especialista altamente calificado para minimizar el error de diagnóstico, que se encuentra entre un 30 % o 40 % dependiendo de la experiencia del profesional que este evaluando [Amin M, 2015], lo que en consecuencia significa que solo hay un 70 % de factibilidad por parte del experto, sin tener en cuenta el desgaste físico y mental que requiere esta actividad, sometiendo a los pacientes siempre a un margen de error considerable.

El error mencionado puede verse reducido al realizar un análisis de imágenes automático, debido a que se disminuye la subjetividad inter e intra personal, en tanto dichas herramientas computarizadas no sufren las consecuencias del esfuerzo físico asociado a la tarea y pueden ofrecer un apoyo a la persona encargada de realizar el diagnóstico. Las imágenes hematológicas son una importante herramienta médica debido a que estas se componen de píxeles, colores e información con patrones difíciles de reconocer para el ojo humano.

3.3. Composición de las imágenes

Los valores de los píxeles conforman una matriz de la cual se puede extraer información de manera matemática ya que adquiere las mismas propiedades operativas de este formato. Así mismo, al agrupar matrices con diferentes características se pueden definir diferentes espacios de representación de una misma imagen, como por ejemplo el RGB que se conforma de tres matrices, rojo (R), verde (G) y azul (B), cuya combinación genera una forma de visualización definida por la percepción de color; en la figura **3-4** se evidencia como los píxeles conforman una imagen.

Las imágenes pueden ser procesadas y analizadas con base en la interacción que existe entre píxeles y la profundidad que tengan, donde la profundidad se refiere al valor máximo que puede adquirir cada píxel. Un ejemplo de esto se encuentra en imágenes a color y blanco y negro, donde en las primeras los píxeles tienen como mínimo 2 bits y las imágenes a blanco

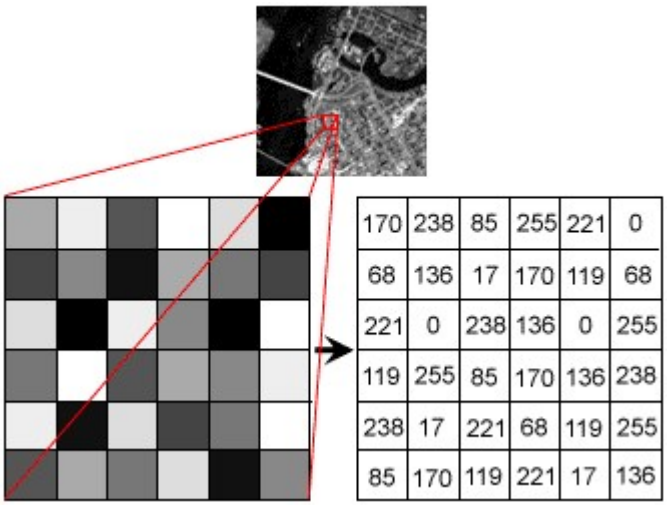

Figura 3-4: Matriz de píxeles que componen una imagen en escala de grises [Morales, 2017]

y negro solo tienen un bit. Entre mayor sea la cantidad de bits de un píxel mas puede ser la cantidad de intensidades que puede representar en una escala de dos (2) elevado a la N (2^N), con N siendo el numero de bits, como se ve en la imagen **3-5**.

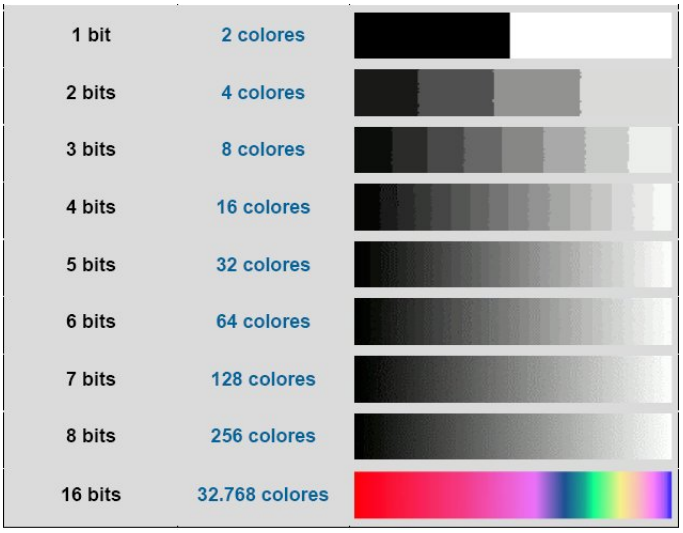

Figura 3-5: Capacidad de un píxel con base a la cantidad de píxeles

3.4. Procesamiento de imágenes

Es todo sistema donde la entrada y la salida son imágenes se presentan una serie de cambios a la entrada para poder extraer o resaltar información por medio de diferentes técnicas de

procesamiento [R., 2013]. A esta serie de acciones se le denomina procesamiento de imágenes y normalmente se realiza en cuatro grandes fases las cuales son: adquisición de imágenes, pre-procesamiento, segmentación y clasificación o selección.

La adquisición de las imágenes es el proceso en el cual se selecciona y se dan parámetros para la toma de la imagen, asegurando las características base del pre-procesamiento y el procesamiento [Jiménez, 2008].

En el pre-procesamiento se realizan adecuaciones de la imagen para que el posterior tratamiento y análisis sea más sencillo o se pueda extraer alguna característica de la imagen en particular, usualmente en esta fase se emplean procesos de modificación de brillo, normalización, ecualización o reducción de ruido como ve en la figura **3-6**.

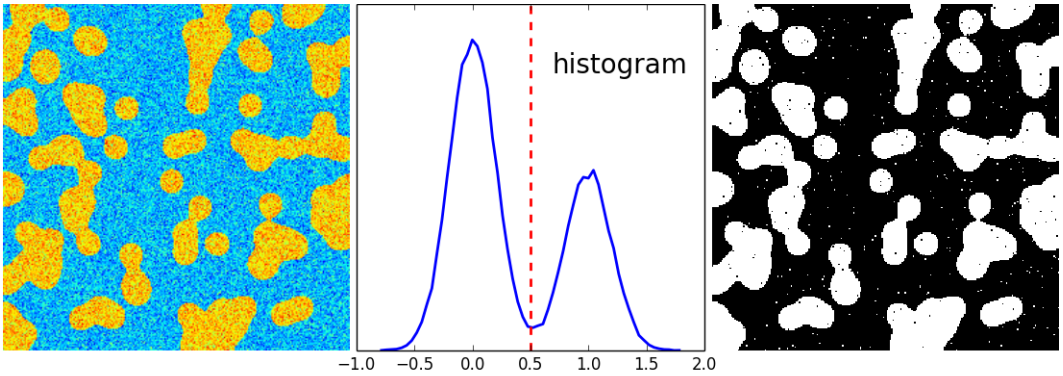

Figura 3-6: Preprocesamiento de una imagen para eliminar ruido por medio del histograma [Mathworks, 2016]

Todo lo anterior se realiza con el fin de preparar la imagen para su posterior análisis, reduciendo ruido, resaltando detalles, eliminando objetos no objetivo, etc. [Miguel A. Castillo Martínez, 2016].

La segmentación consiste en separar los objetos de interés sin perder los detalles del mismo [Miguel A. Castillo Martínez, 2016], este proceso se puede realizar por medio de la detección de bordes, binarizando, por watershed u otras funciones como se ve en la figura **3-7**.

Una vez realizadas las tareas de pre-procesamiento y segmentación se requiere realizar en muchos casos operaciones con filtros morfológicos para limpiar aun mas la imagen y eliminar cualquier tipo de objeto que no sea de interés.

Al proceso de limpieza de una imagen por medio de filtros morfológico se llama proce-

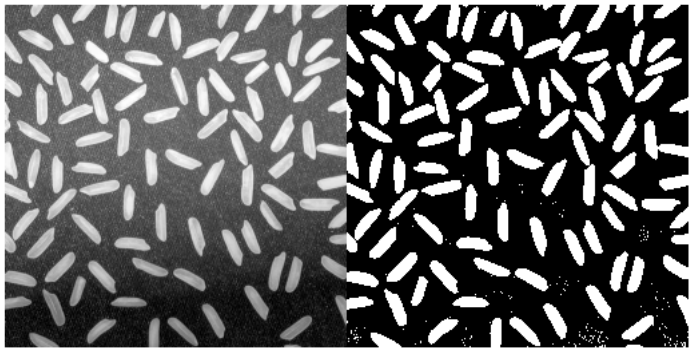

Figura 3-7: Imagen binarizada por metodo de segementacion, para ampliar detalles [Mathworks, 2016]

samiento morfológico y se enfoca en la extracción de figuras básicas (círculos, cuadrados, diamantes, hexágonos, etc.) con diferentes tamaños para poder realzar o disminuir la geometría de los objetos internos de las imágenes.

Al aplicarse los filtros morfológicos se expanden las formas, como se muestra en la figura **3-8**, en la cual se tienen puntos que se exageran al ser sometidos a un filtro morfológico de círculo, esto mismo funciona al reducir un objeto de la misma manera, permitiendo el efecto contrario al mostrar los objetos separados.

3.5. Tipos de detección de objetos en imágenes

A lo largo del tiempo se ha trabajado en formas de automatizar el diagnóstico centrándose en características visuales del núcleo y del citoplasma, siendo el progreso de uno reflejo del otro, es decir, que las dos variables están relacionadas en el diagnóstico de LLA, lo que dio como resultado automatizaciones centradas en esas dos características principales [Amin M, 2015] centrándose también en el tamaño del linfocito, la redondez del núcleo y su relación con el citoplasma.

Las características mencionadas sirven para la diferenciación de linfocitos, además se complementan con características de los objetos dentro de las imágenes como detalles morfológicos, de contraste, bordes y excentricidad [Vaguela, 2015, Scotii, 2004, Putzu L, 2014]. Las anteriores son usadas como parámetros para aislar y ubicar los objetivos dentro de la imagen [Jiménez, 2008], también hay otras características que se pueden extraer como son: el eje superior, el eje inferior, orientación, entre otros [Putzu L, 2014].

Figura 3-8: Puntos sometidos a filtro morfológico de dilatación con círculos

Los descriptores de los objetos son parámetros que pueden ser diferenciales para las herramientas de clasificación automatizadas para definir una clase en un objeto especifico, como lo son en este caso los linfocitos.

3.6. Métodos de clasificación

Los linfocitos y las propiedades de imagen que contienen se pueden someter a clasificadores automatizados, lo cual sirve para resolver problemas de clasificación referido como la acción de decisión que se realiza para determinar la clase de algo; existen problemas binarios o de multi-clasificación [Enrique J, 2014] estos emplean métodos matemáticos robustos para lograr su objetivo y se han desarrollado varios tipos de clasificadores tales como maquinas de soporte vectorial [Putzu L, 2014], redes neuronales [Jagadeesh, 2013], C-means y K-means [Dr.Karthikeyan, 2017] y Random Forest [Mishra, 2017] que han sido usados para la clasificación de linfocitos, por medio de las características de referencia usadas en el aprendizaje de cada algoritmo, en la figura **3-9** se ve representado como funciona la clasificación normalmente.

El aprendizaje interno automatizado de un software es el proceso en el cual se definen los datos de la clase a clasificar [Morales, 2017], este aprendizaje matemático da como respuesta un tipo de equivalencias matemáticas internas que permiten al software determinar con base en su aprendizaje que tipo de objeto es el de entrada.

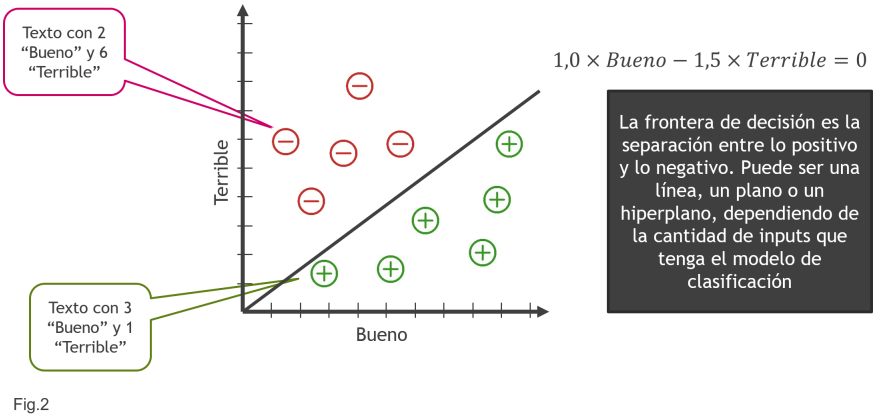

Figura 3-9: Representación de como funciona una clasificación y como esta cambia dependiendo de la cantidad de iputs

3.6.1. Random Forest

Este método de predicción usa múltiples clasificadores o regresores (siendo estos los árboles del bosque) para poder determinar a que clase pertenecen las entradas del sistema. EL método disminuye la correlación entre los árboles, por medio de la aleatoriedad que tiene el sistema y entrega como resultado o predicción un promedio de las clasificaciones que realizó [Campos, 2017].

El algoritmo consta de dos partes principales las cuales son: crear los árboles de regresión y la introducción de datos de manera aleatoria siendo N la cantidad de datos y la cantidad de pruebas realizadas por el sistema [Biau, 2012].

Aunque el método funcione correctamente se tiene como riesgo el sobre-ajuste, dependiendo de la cantidad de árboles realizados y sus valores individuales, los cuales si llega a ser una cantidad excesiva se puede demorar mas de lo requerido y si son pocos pueden llegar a clasificar erróneamente.

3.6.2. Suport vector machines- SVM

Este método de clasificación emplea hiperplanos entre las clases de entrada, intentando diferenciar una de otra matemáticamente y buscando el mejor plano entre las entradas para su futura clasificación [H, 2017].

Utiliza rectas, superficies, planos o hiperplanos para poder realizar la diferenciación dependiendo de la cantidad de clases de entrada ingresada, estos siempre por naturaleza del sistema, tendrá una máxima distancia entre las clases y también, entre el plano y los datos,

para generar el hiper plano se realiza por medio de la siguiente formula matemática

$$\omega^T \times X = 0 \tag{3-1}$$

Donde es otra forma de expresar una recta, pero esta formula cambia cuando se requiere separar dos clases donde

$$\omega^T \times X \geq 1 \tag{3-2}$$

$$\omega^T \times X \leq -1 \tag{3-3}$$

Para una clase debe ser mayor a uno mientras que para otra debe ser menor a uno, indicando que hay dos hiper planos posible, uno para cada clase, estos cambian dependiendo del los valores de ω y de X que son los valores de peso que se emplea al sistema, y una vez se tengan los dos hiper planos, se genera un tercero y definitivo entre estos que es el que maximiza la distancia entre los dos anteriores. [Brummel, 2002]

Las SVM pueden ser configuradas para evitar el sobre ajuste sin importar la cantidad de clases de entrada, esto se realiza por medio del uso de kernel en el sistema y la configuración de estos, que permite hacer el sistema mas rígido o flexible al momento de crear los diferenciadores [Mishra, 2017].

La configuración anterior puede verse alterada dependiendo de la forma en que los datos estén distribuidos cambiando el tipo de clasificación a realizar, por ejemplo:

- Clasificación binaria separable: este tipo de clasificación se da cuando los valores a clasificar presentan una separabilidad clara y es posible separarlas con una linea recta en la mayoría de los casos, este tipo de configuración solo se puede lograr cuando existen solo dos clases

- Clasificación binaria cuasi-separable: se presenta cuando los valores de entrada tienen una separación clara para el usuario pero la maquina no puede realizar una separación directa, para estos casos se configura de tal manera que el sistema omita algunos datos para la realización del plano, logrando una separación completa sin tener en cuenta los valores mas cercanos entre ellos permitiendo un plano recto entre clases.

- Clasificación no separable linealmente: es cuando una clase rodea a otra imposibilitando la generación de un plano recto, para estos casos se emplean kernel no lineales que permitan separar las clases con un plano curvo o hiperbólico, de esta manera se puede rodear una sola clase aislándola de otra y formando una clasificación circular en algunos casos.

Todas las configuraciones anteriores permiten clasificar diferentes tipos de clases y se puede extrapolar a varias clases con los mismos principios, variando los valores matemáticos de la inteligencia artificial.

Por ejemplo, en un método de clasificación se disminuye el tiempo de procesamiento por imagen a 1001us (milisegundos) mejorando el tiempo de procesamiento de cada una, aunque este sistema tenga que ser supervisado por un oncólogo especializado, permite adquirir una calidad similar a la del trabajo de los profesionales especializados [Piuri, 2004].

4 Metodología y Resultados

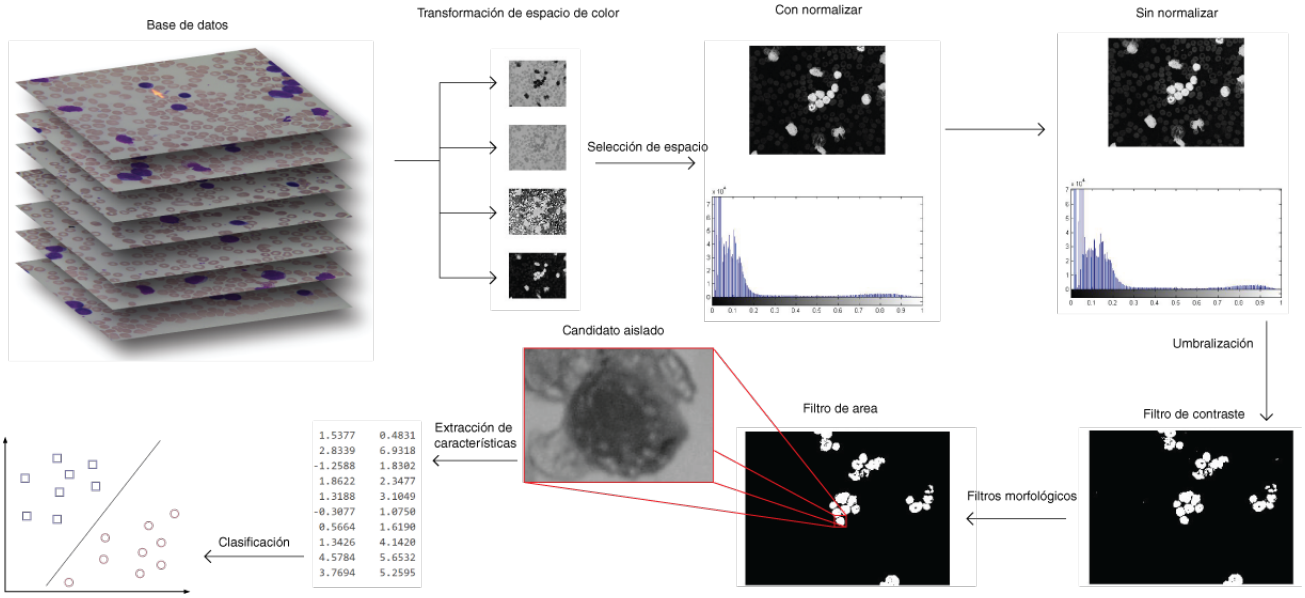

Figura 4-1: Fases de desarrollo del proyecto de grado

El proceso de desarrollo del método propuesto para la clasificación de linfocitos afectados por leucemia linfoblástica aguda en imágenes hematológicas, contempla distintas fases como se evidencia en la figura **4-1**. Este, inicia con la obtención y caracterización de las imágenes, la descripción de la base de datos y la asignación de etiquetas de diagnóstico. Posteriormente, se efectúa una tarea de pre-procesamiento para selección de color, homogeneización de intensidades, la corrección de ruido, seguido de tareas de segmentación que permiten obtener una imagen binaria preliminar que resalta los posibles candidatos de leucocitos.

Después de lo anterior, se extraen características de las imágenes preprocesadas para determinar a qué clase pertenece cada elemento de interés, obtenido en el proceso de segmentación. Una vez extraídos, estos descriptores son organizados en bloques que llamaremos diccionarios, los cuales serán evaluados dentro de una máquina de soporte vectorial (SVM) que permite tener una clasificación de los objetos y una confirmación de viabilidad por medio de una curva de características operativas del receptor (ROC por sus siglas en ingles).

A continuación cada uno de los procesos mencionados será profundizado, ofreciendo detalles metodológicos del desarrollo de cada una de las fases.

4.1. Base de datos

Teniendo en cuenta las consecuencias temporales del acceso a pacientes (con LLA y control) para la extracción de muestras sanguíneas y la gestión ética del manejo de datos de los mismos, se prefirió trabajar con bases de datos con imágenes previamente adquiridas que sirvieran de referencia para establecer un control sobre los datos de entrada y el resultado obtenido. Además, de tener un banco de información ya indexada y diagnosticada, también permite conocer las características del objeto sobre el cual se está trabajando, logrando así un ambiente controlado para obtener pruebas y resultados confiables.

Así pues, los datos usados en este trabajo se extrajeron de la Base de Datos de Imágenes de Leucemia Linfoblástica Aguda denominada ALL-IBD (por sus siglas en inglés), que se centra en imágenes hematológicas de linfocitos con y sin leucemia para el desarrollo de algoritmos de segmentación y clasificación entre los dos grupos de diagnóstico mencionados [Scotii, 2004]. Esta batería de imágenes dispone de etiquetas con diagnósticos efectuados por un grupo de profesionales en oncología, lo cual proporciona un *ground truth* que garantiza la validación de la metodología propuesta [Scotti, 2006].

ALL-IBD está dividida en dos grandes bancos de imágenes, el primero (ALL-IDB1), está destinado a evaluar los algoritmos de segmentación y clasificación, mientras que la otra (ALL-IDB2), está construida para evaluar el procedimiento de clasificación. En la tabla **4-1** se muestra un resumen del contenido de la base de datos, en la que se específica la cantidad de imágenes, la resolución de las mismas, el número de linfocitos candidatos a ser diagnosticados y la totalidad de objetos (elementos) en general que comprenden la imagen (glóbulos rojos, glóbulos blancos, artefactos, etc).

La ALL-IDB1 contiene imágenes de frotis sanguíneo obtenidas por microscopía óptica, es decir, se pueden observar todos los elementos que un especialista vería en su ejercicio de diagnóstico. Dentro de esta base de datos se encuentran muestras afectadas por leucemia (49) y muestras sanas (59), sumando un total de 108 muestras de visualización completa, dentro de estas imágenes se puede contemplar la diferencia directa entre una muestra sana y una muestra enferma como se observa en las figuras **4-2** y **4-3** [Labaty, 2011]. Otro aspecto para aclarar, es que los elementos en la tabla hacen referencia a la cantidad de objetos (linfocitos candidatos) que se pueden identificar en la base de datos entre glóbulos rojos, blancos y agentes coagulantes.

Características de la adquisición de la imagen		
Cámara Canon PowerShot G5 Aumento del microscopio de 300 a 500 Formato de imagen JPG profundidad de color de 24bg		
Parámetro	ALL-IBD1	ALL-IBD2
imágenes	109	260
Resolución	2592x1944	257x257
Elementos	39000	260
Linfocitos candidatos	510	130

Tabla 4-1: Características de las bases de datos de imágenes adquiridas [Scotti, 2006]

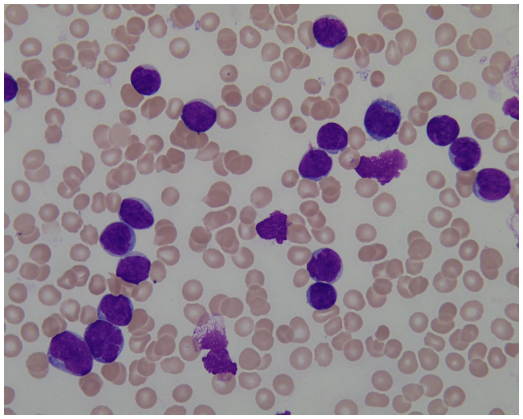

Figura 4-2: Muestra de frotís de sangre afectada por LLA

La segunda parte de la base de datos se compone de segmentos de imagen (parches), cada uno conteniendo un linfocito proveniente de las imágenes de la ALL-IDB1 como parches de imagen. Con esta se espera que se realicen pruebas de clasificación diferencial del estadio de LLA (L1, L2 y L3), sin embargo, esta tarea de determinación entre varias clases no se realizará dentro de este trabajo de grado.

La base de datos escogida solo proporciona la información binaria de positivo o negativo para LLA, codificada con 1(uno) o con 0 (cero) respectivamente, Esto se evidencia dentro de la figura **4-4**. Debido a este tipo de marcación no se hacen distinciones acerca de las características de cada linfocito.

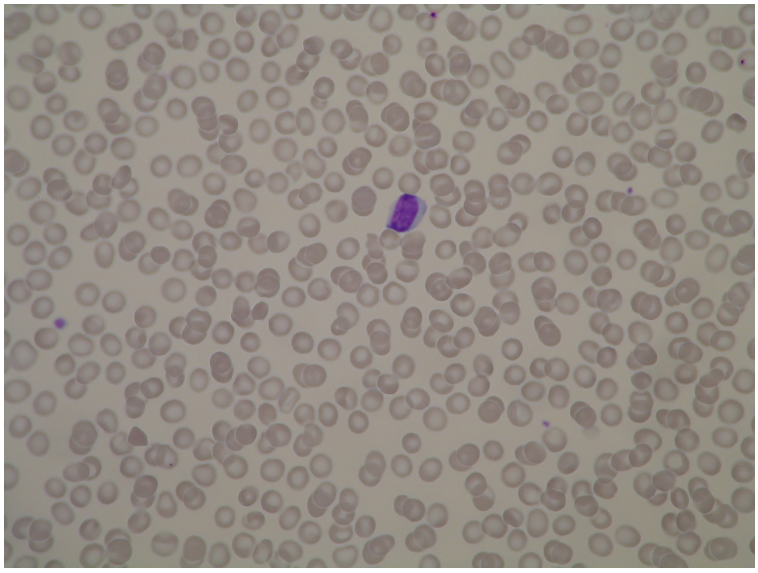

Figura 4-3: Muestra de frotís de sangre sana

Se escogió esta base de datos debido a la facilidad de adquisición que propone en términos comparativos con otras bases de datos, además de entregar imágenes de linfocitos con morfología común y poco común, lo que permite entrenar los algoritmos para poder diferenciar distintos tipos de morfología permitiendo adquirir mejores resultados en los estudios realizados por los autores [Scotii, 2004].

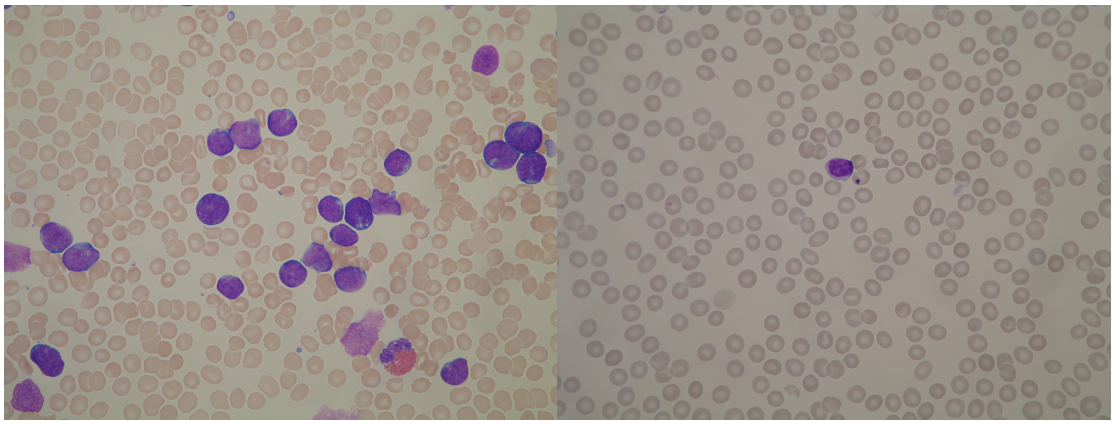

Figura 4-4: Contraste de muestra positiva y negativa de LLA [Scotti, 2006]

Además, con la base de datos referenciada se han realizado diferentes trabajos que han presentado buenos resultados bajo diferentes métodos y con diferentes propósitos, como lo es el análisis morfológico de linfocitos en el trabajo de F.Scottii [Scottii, 2005].

4.2. Espacio de color

Una vez seleccionada la base de datos se procede a preprocesar las imágenes incluyendo la determinación del espacio de color sobre el cual se trabajarán las imágenes. Para esto se realiza un estudio y una comparación acerca de los diferentes espacios de color junto con sus ventajas y desventajas de los mismos.

Al estudiar los espacios de color en los cuales se puede resaltar los linfocitos, por medio del estudio de los diagramas de dispersión se evidencia que en la gráfica de dispersión por canales del HSV (figura **4-6**) en contraste con la gráfica de dispersión de canales del RGB (figura **4-7**) los canales son mas separados unos de otros, lo que permite hacer una mejor separación en el espacio de color de HSV [Ariffin, 2012], teniendo en cuenta lo anterior como se aprecia en la figura **4-5** el canal de saturación del HSV permite trabajar los linfocitos con ruido disminuido desde la selección del mismo.

En el grupo de imágenes de la figura **4-5** se puede realizar un contraste con la matriz de contraste (F) o la de brillo (H) estas se pueden usar como máscaras, sin embargo, no eliminan el ruido de forma constante, contrario a la matriz de saturación. En la imagen F se observa que hay ruido de fondo que puede interferir en el procesamiento y, en la imagen H, no se puede realizar con seguridad una distinción completa entre los candidatos positivos y los negativos, por el contrario, en la matriz de saturación se facilita la distinción de lo que es y lo que no es linfocito.

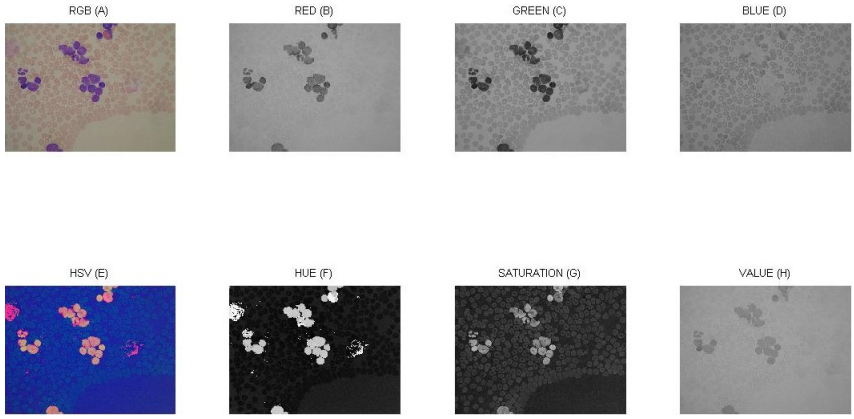

Figura 4-5: Componentes de RGB y HSV, a la izquierda se encuentran las imágenes originales y se desglosan en cada una de sus componentes hacia la derecha

En los componentes de RGB se puede usar la matriz de composición de verdes, debido a que se distinguen los linfocitos, pero al encontrar errores consecuentes al ruido, fue descartada.

Al escoger un canal que permitiera la distinción de linfocitos y la máxima disminución

de ruido posible, se comenzó a trabajar sobre ese canal como un punto de inicio para el procesamiento de la imagen, lo cual se hace todo en valores de 1 a 255 como si se trabajase en una matriz de tonos de grises, como se evidencia en la figura **4-5**, en la cual se ven los canales como si fueran grises. Se hizo un estudio por medio de diagramas de dispersión para evidenciar el espacio que permite resaltar los objetos de interés con mayor facilidad, tanto visual como matemáticamente. El espacio de color RGB y el HSV tienen diferencias marcadas en los diagramas de dispersión como se puede ver en las figuras **4-6** y **4-7**

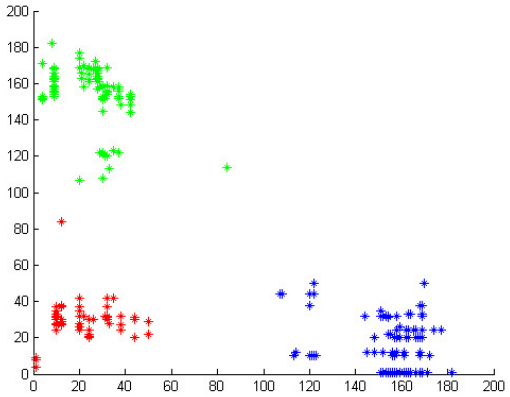

Figura 4-6: Representación de datos espacios de imágenes de linfocitos en HSV es decir matiz (rojo),saturación (verde) y brillo (azul) respectivamente

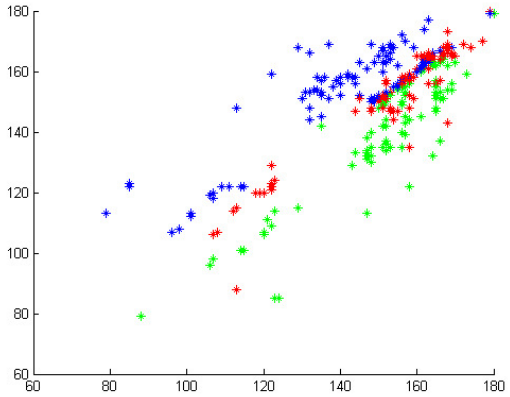

Figura 4-7: Representación de datos espacios de imágenes de linfocitos en RGB, siendo cada espacio de color representados por R en rojo, G en verde y B en azul

Al contrastar las figuras **4-6** y **4-7**, es evidente que el HSV tiene una mejor dispersión de datos, lo que hace que sea una opción viable para comenzar el pre-procesamiento, al observar el histograma y la imagen de cada canal (figura **4-8**) se puede ver que la dispersión de datos

matemáticamente también es mejor, disminuyendo el ruido a datos menores en comparación con los datos de los objetos de interés.

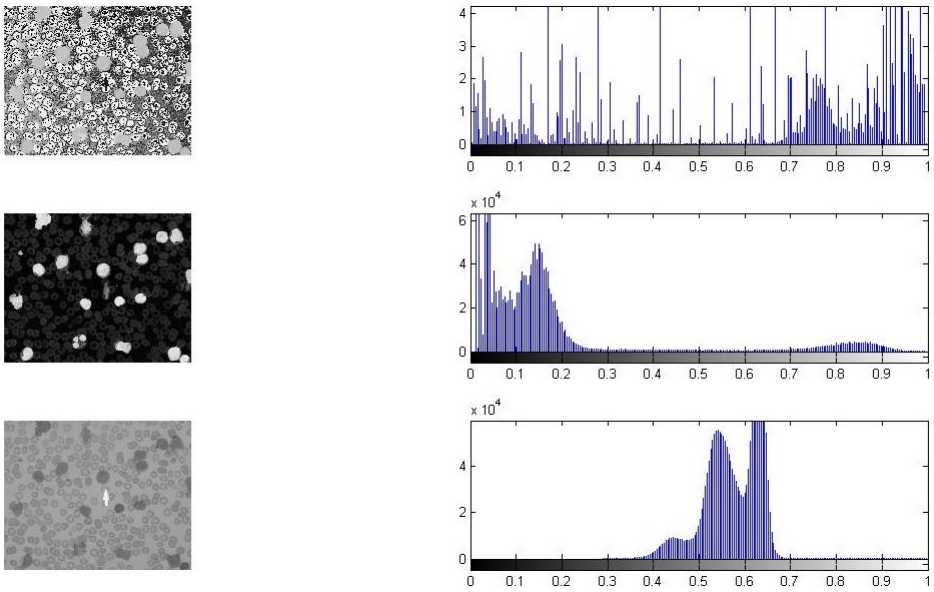

Figura 4-8: Conversión de una imagen HVS con sus respectivos histogramas

4.3. Normalización

Con las imágenes filtradas en el espacio de color seleccionado se procede a realizar una normalización que permita disminuir el efecto de variabilidad de los datos por el posible origen de las mismas, considerando tanto las variaciones de iluminación como las de tono en algunas de las imágenes.

Al realizar un análisis del histograma de la imagen resultante, se extrajeron valores para la eliminación de ruido en las imágenes a escala de grises, como se ve en la figura **4-9**.

Con el análisis del histograma se decidió realizar una normalización logarítmica debido a que esta permite resaltar el contorno de los objetos menos visibles y poner la intensidad de las imágenes sobre la misma escala. Para esto, la normalización hace uso de la ecuación 4-1:

$$s = T(r) = \sum_{0}^{r} P_r(w)\mathrm{d}w \tag{4-1}$$

En la ecuación 4-1, $T(r)$ es la función de transformación, $P(r)$ es la función de densidad de probabilidad de la imagen, w es la variable de integración de la imagen y, r es la posición

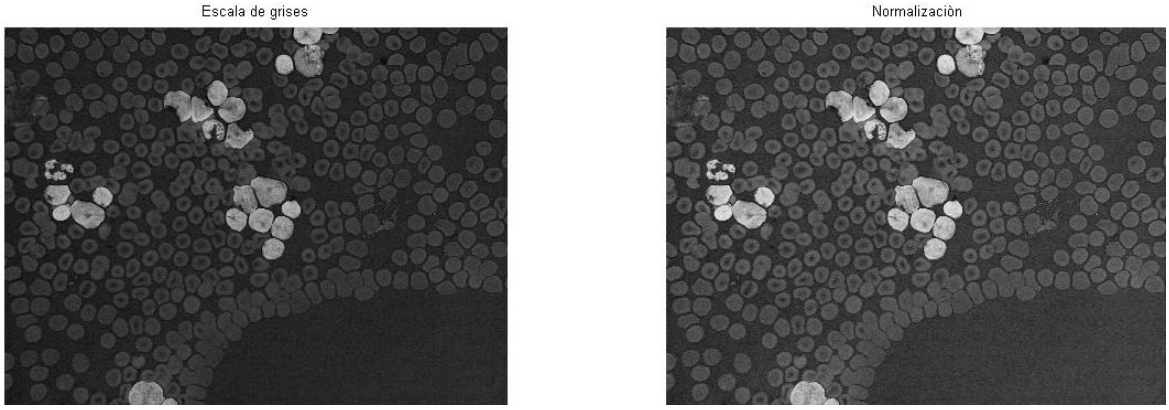

Figura 4-9: Comparación imagen escala de grises con imagen normalizada

del píxel en la imagen, lo que hace la función anterior es pasar de un gráfica lineal a una logarítmica acentuando los valores cercanos a cero y atenuando los valores cercanos a uno, evitando sesgos matemáticos y manejando un rango de valores menor al natural de la imagen [González, 2008].

Esta normalización permite contrastar las secciones oscuras de la imagen tornándolas más claras, lo que permite notar mejor los bordes de los objetos internos y definir a qué pertenece a cada objeto [Jagadeesh, 2013], de igual manera, los datos internos de la imagen se escalan para que siempre queden dentro del mismo rango matemático, lo que da como resultado mejor distribución de los datos, lo anterior, se puede evidenciar en la figura **4-10**.

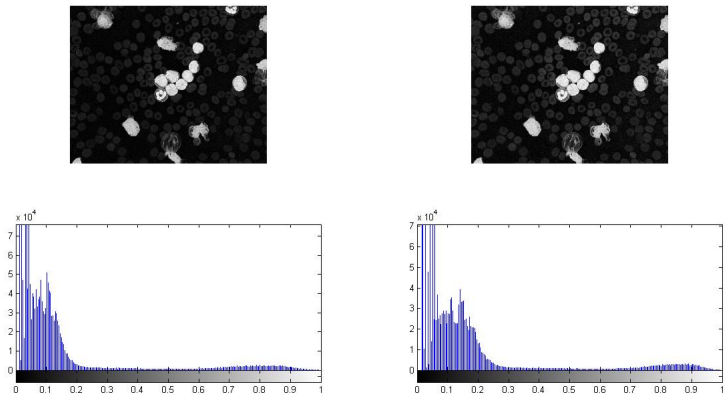

Figura 4-10: Muestra de imagen normal (izquierda), imagen normalizada (derecha)

4.4. Umbralización

Se realiza una umbralización haciendo uso del método de OTSU, lo que genera una discriminación completa del fondo y los objetos de interés.

Una vez normalizada la imagen se somete al proceso de umbralización OTSU lo que adquiere valores binarios y elimina los objetos que no son de interés, permitiendo centrar el estudio en los candidatos a linfocitos, ver figura **4-11**.

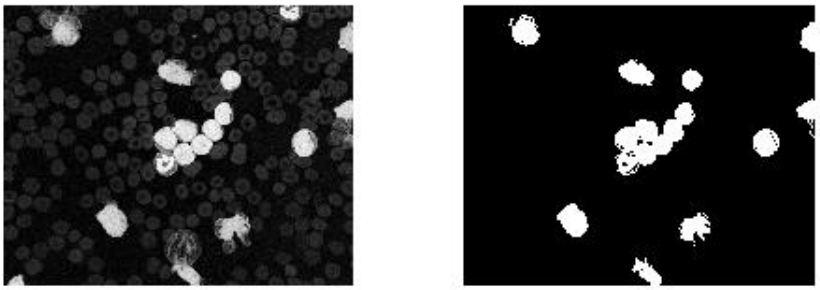

Figura 4-11: Imagen normalizada (izquierda), Imagen umbral izada (Derecha)

Debido a que OTSU es una umbralización dinámica entre las clases dominantes del histograma, esta se empleó para generalizar los datos de entrada de las imágenes, como se aprecia en las figuras **4-11** y **4-12**, permitiendo transparencia en el método de adquisición de la imágenes; en las siguientes ecuaciones se explica como funciona este método:

$$\alpha^2 = \omega_B(\mu_B - \mu)^2 + \omega_F(\mu_F - \mu)^2 \tag{4-2}$$

$$\omega = \sum_{n=1}^{N} P(n) \tag{4-3}$$

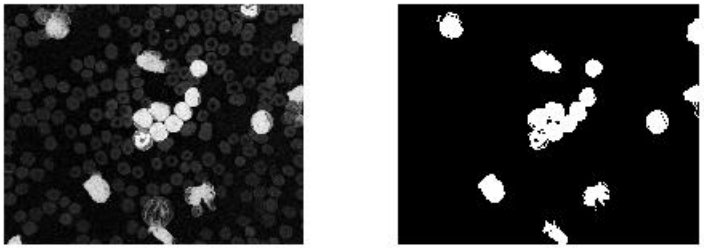

Figura 4-12: Muestra visual y matemática de la umbralización, donde se evidencia como se han perdido objetos a raíz del proceso de umbralización

B = Fondo, F = Objetivo , n= posición del píxel, N= Cantidad máxima de píxeles de la clase a la que pertenece

Donde, el valor máximo de α^2 es el umbral óptimo para la imagen dada, ω es la probabilidad acumulada de que suceda una clase sobre otra, μ es la media de una clase u otra (siendo b el fondo y f el objetivo) y la media común de toda la imagen, de esta manera el sistema puede discriminar el fondo y el objetivo.

Se observó que los objetos aislados al final corresponden a los candidatos de la imagen original, ver figura **4-13**, aunque siguen existiendo remanentes de ruido y poca distancia entre los objetos, por lo que se continúa con los filtros morfológicos para eliminar ambas variables y tener objetos aislados.

4.5. Segmentación y extracción

Una vez adquiridas las imágenes umbralizadas en el canal de saturación (HSV) se aislaron los componentes de interés, que en este caso son los linfocitos, para lo cual existen diferentes tipos de técnicas de segmentación como lo son, watersheet [Mishra, 2017], Clus-

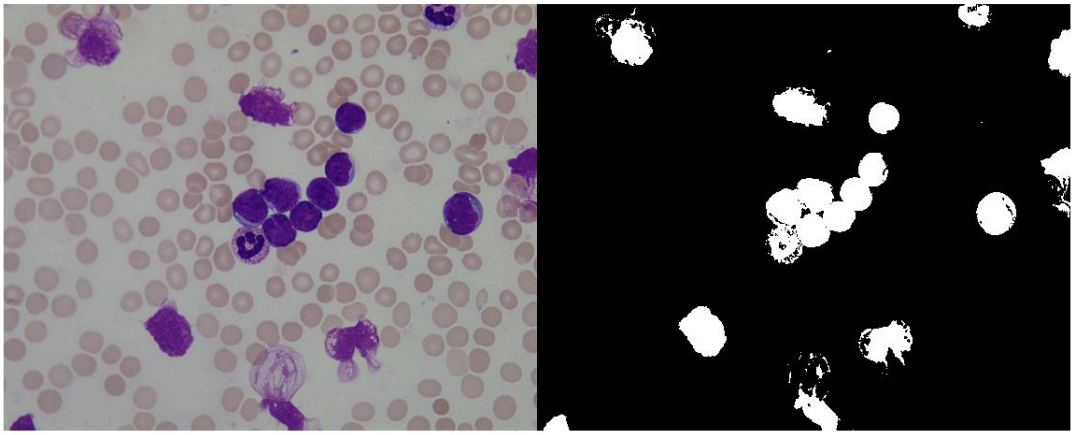

Figura 4-13: Representación del resultado de la normalización junto con la binarización de OTSU, Izquierda: linfocitos, Derecha: linfocitos aislados por el filtro

tering [Srisukkhama, 2017], [Jiménez, 2008], K-means [Kandil, 2016] [Sarrafzadeh, 2015], C-means [Dr.Karthikeyan, 2017], todas ellas realizadas en función de las características de interés, como la forma o la circularidad y el tamaño, entre otros factores.

Para refinar la detección de posibles candidatos se usaron filtros morfológicos aplicados en los objetos obtenidos tras umbralizar, separando posibles objetos unidos, oclusiones o remanentes de ruido **4-16**. Posteriormente a partir de las coordenadas de sus centroides se extraen parches que contienen los mismos con el objetivo de extraer características que permitan discriminar si dicho objeto perteneces a la clase "linfocito".

4.5.1. Filtros morfológicos

Como etapa previa al filtrado de forma, se extraen los bordes de la imagen binarizada para definir los limites de los objetos encontrados, mediante el uso del método de Roberts, ver figura **4-14** [Departamento de Ingeniería Electrónica, 2005], para la aproximación de las derivadas de las intensidades en los ejes vertical y horizontal. Posteriormente, al resultado encontrado se le aplicó una **dilatación morfológica** en la imagen, buscando rellenar posibles vacíos en el borde de la misma con el fin de obtener contornos cerrados para cada candidato.

Para la mencionada tarea de filtrado morfológico se usaron diferentes kernels (formas) incluyendo discos, diamantes y cuadrados con diámetros de 8 y 12 píxeles, que responden a las formas que toman los linfocitos en las imágenes adquiridas, como se evidencia en la figura **4-15**.
Luego se usaron estos mismos elementos filtrantes para aplicar una **erosión morfológica**

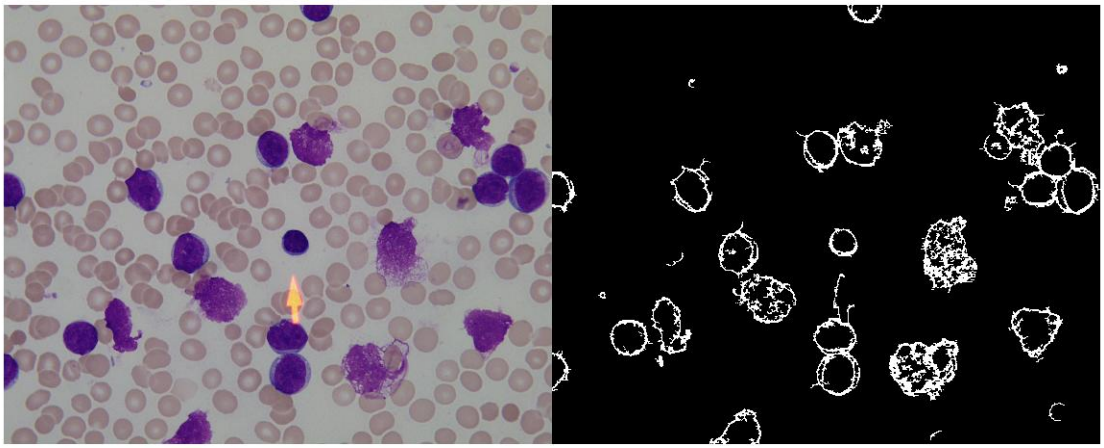

Figura 4-14: Izquierda: imagen original. Derecha: Bordes de los objetos encontrados, se puede notar que se perciben objetos ajenos a linfocitos

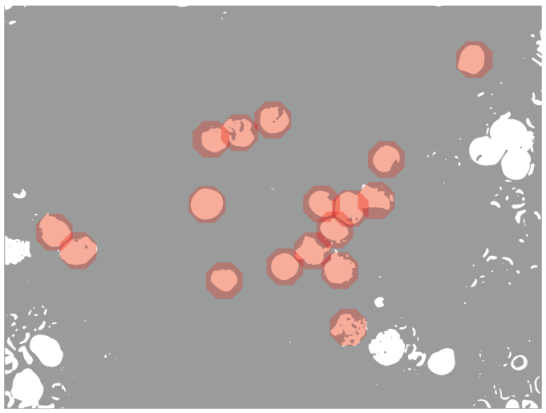

Figura 4-15: Formas de filtros morfológicos básicas

en busca de separar los posibles objetos y eliminar elementos pequeños así como el ruido en la imagen. Es así que se obtiene una versión depurada de los posibles candidatos como se evidencia en la figura **4-16**.

4.5.2. Separación o extracción de candidatos

Todo el proceso descrito en la subsección anterior (filtrado morfológico) se realizó con el fin de identificar un grupo de candidatos a los que se les pudiera extraer diferentes características destinadas a discriminar la clase a la que dicho objeto pertenece (linfocito, no linfocito).

Posterior a esta tarea, se usaron los centroides como ubicación espacial dentro de la ima-

gen, que permita localizar y extraer los candidatos, realizando un marco sobre el objeto y aislándolo en una imagen nueva, tal y como se evidencia en la figura **4-17** C.

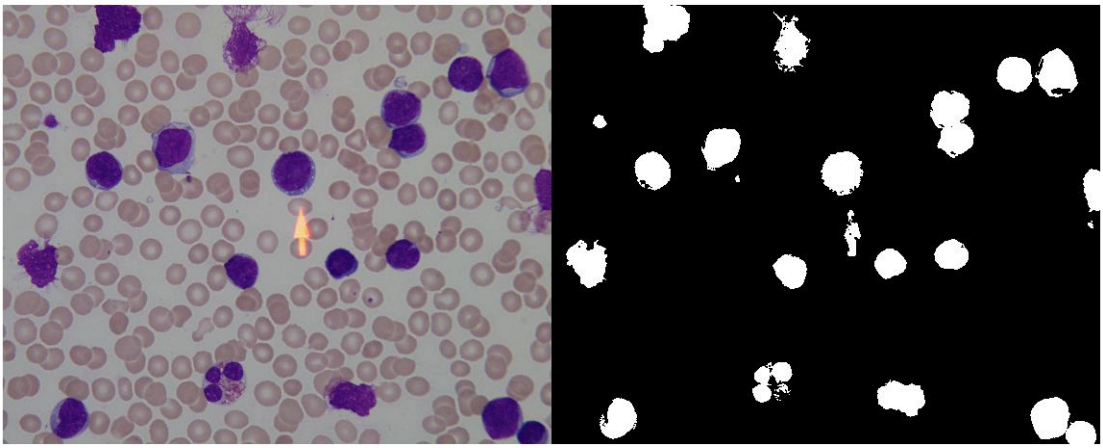

Figura 4-16: Izquierda: imagen original en espacio RGB, Derecha: imagen resultante tras aplicar los filtros morfológicos

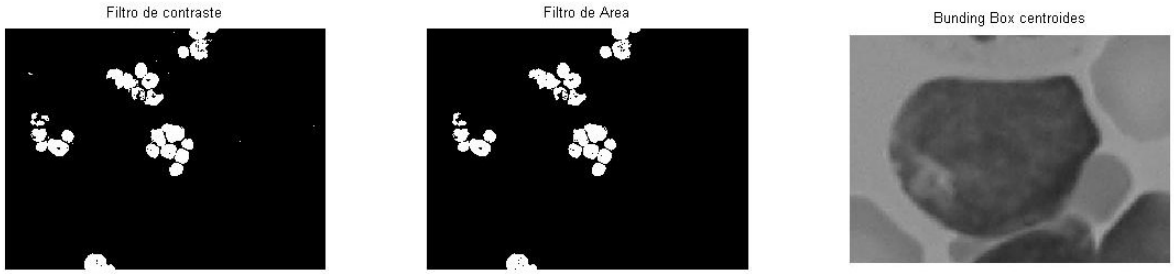

Figura 4-17: Proceso de segmentación de izquierda a derecha, A: filtro de contraste, B: filtro de área y morfológicos, C: Resultado linfocito aislado

Para la adquisición del centroide se toma el centro de masa de un plano, de esta manera se adquieren los ejes de X y Y máximos, donde la intersección de ellos es la característica a analizar, en la siguiente ecuación se muestra como se calcula el centro de masa de un objeto irregular.

$$\overline{X} = \frac{\sum_a^b X[f(x) - G(x)]\mathrm{d}x}{A} \qquad \overline{Y} = \frac{\frac{1}{2}\sum_a^b X[f(x)^2 - G(x)^2]\mathrm{d}x}{A} \tag{4-4}$$

Donde para este caso las funciones $f(x)$ y $G(x)$ se toman como funciones elipsoides.

Al final se adquieren Linfocitos aislados de la siguiente manera **4-18**:
A pesar que la separación de linfocitos es óptima no tiene inmunidad al ruido al momento

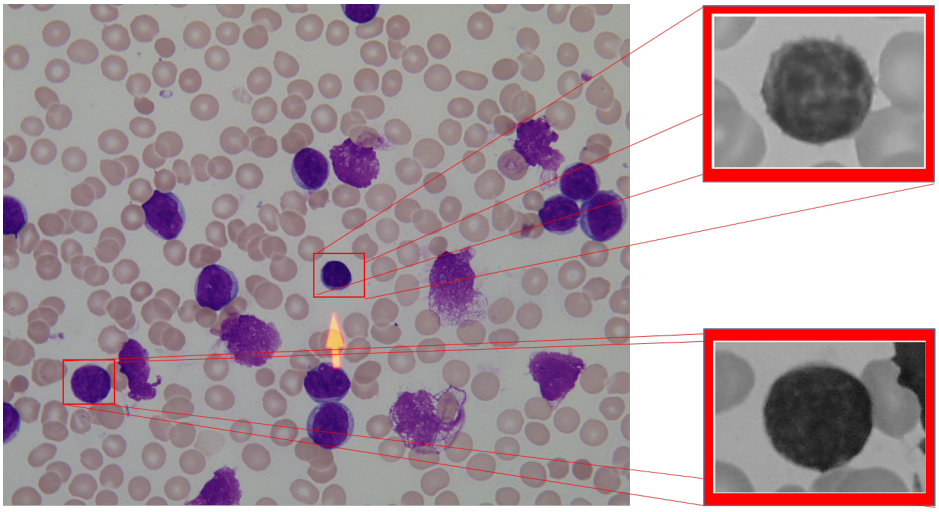

Figura 4-18: Representación de como se aíslan los linfocitos

de tener imágenes con muchos linfocitos unidos, donde los separa pero las características son difusas debido a que esta unido con uno o mas linfocitos, como se ve en la figura **4-19**. Finalmente, en busca de definir el tamaño del marco de extracción de candidatos (parche) se realizó una comparación entre todos los posibles objetos buscando obtener la máxima medida de esta característica asociada, encontrando que el valor de tamaño definido debe ser de 250×250 píxeles. Es importante que todos los parches tengan el mismo tamaño para que se homogenice el proceso automatizado de extracción de características, sin importar el tamaño del candidato.

La extracción de características esta vinculada con el proceso de extracción de linfocitos donde los linfocitos aislados tanto positivos como negativos muestran características discriminadoras para el usuario. Al observar las figuras **4-20** como aislado negativo y **4-21** como aislado positivo se evidencia diferencias físicas entre ellas, por lo cual es necesario extraer esas características en valores para el computador.

Las imágenes anteriores el sistema las reconoce como candidatos a LLA con diferencias en sus características, que aunque aun no defina por si mismo si son positivos o negativos, los reconoce como objetos candidatos a ser un linfocitos, es por esa razón que se quiere de un proceso de enseñanza para que el sistema pueda reconocer los positivos y los negativos
Segmentados los objetos de interés, se extrajeron las características de esos objetos (centro, área, perímetro, dirección, redondez, máximo eje X y máximo eje Y) y se organizó de tal manera que fuera fácilmente manejable para su estudio y selección, por lo que se colocó dentro de archivos que serán conocidos como matrices de características.

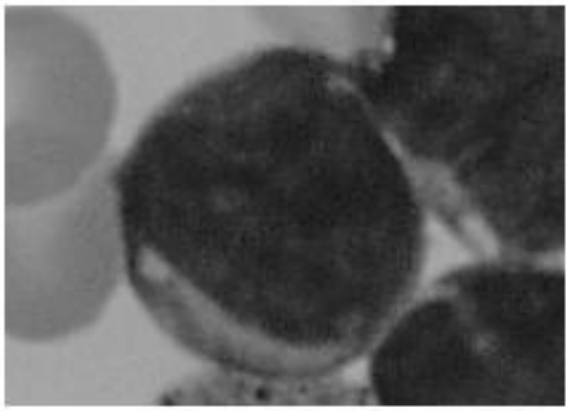

Figura 4-19: Extracción de linfocito unido a otro

4.6. Procesamiento y extracción de características

Durante el proceso de adquisición de linfocitos para luego ser clasificados, se etiquetaron todos los objetos encontrados en la imagen para adquirir las propiedades espaciales que tienen estos, tales como: la ubicación dentro de la imagen, redondez y área total, entre otros.

Una vez se logró ubicar los objetos, se realizó una verificación de posición, ver figura **4-22**, lo que permitió segmentar la imagen en sub-imágenes de linfocitos y se crearon diccionarios de características de los objetos encontrados, los cuales están directamente relacionadas con el diagnóstico manual realizado por el profesional como lo son la excentricidad, el área convexa y la solides.

La excentricidad de los objetos es un indicador de la deformación axial de una circunferencia (si tiene diámetros diferentes), mientras mas redondo sea el objeto no solo asegura que es un linfocito, sino que también ayuda a determinar cuales ya son inmaduros por el tamaño de su núcleo, cosa que de igual forma se analiza en el método manual de manera individual con cada uno, como se ve en la figura **4-23**. Esto se encuentra asociado a características morfológicas descritas por la FAB donde la falta de simetría en los ejes de la circunferencia demuestra la afectación de la célula [Sala, 2003].

El área convexa se refiere al área interna de llenado del objeto dentro de la imagen y se

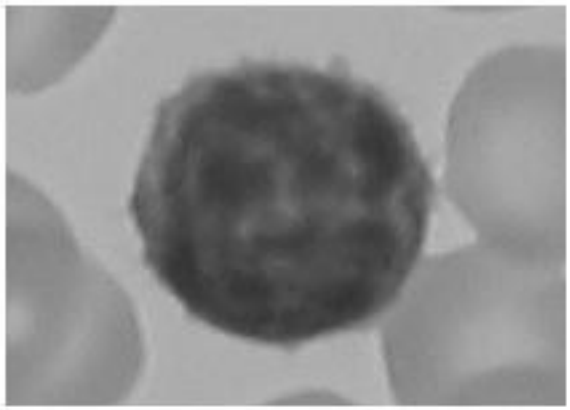

Figura 4-20: Linfocito aislado negativo para LLA

relaciona directamente con el citoplasma interno del objeto, esta característica está vinculada con la relación citoplasma/núcleo que según la FAB indica inmadurez de la célula y es síntoma de leucemia.

La solidez de un objeto se refiere a la textura del mismo, que aunque no sea un factor determinante en el análisis manual, el algoritmo si puede determinar la textura de los objetos, por lo que se optó por usarla como parámetro extra debido a que el núcleo de los linfocitos infectados por leucemia tiene diferente textura respecto de uno sano, debido a que la leucemia cambia el núcleo de las células volviéndolas menos "solidas", o lisas para la visión artificial.

Las características seleccionadas son filtradas por su relación con el método manual que ejecuta el profesional, lo que permite descartar algunas como la orientación, la cual es transparente al diagnóstico, dando como resultado una cantidad de propiedades relacionadas con el método manual, las cuales se explican a continuación:

- Área de relleno: esta característica se usó teniendo en cuenta la relación de citoplasma núcleo que tienen los linfocitos. Al aumentar el tamaño del núcleo el área de llenado interno se aumenta de manera proporcional, incrementando este valor convirtiéndolo en una característica a tener en cuenta.

  Para adquirir este valor el sistema realiza un conteo de los píxeles que conforman

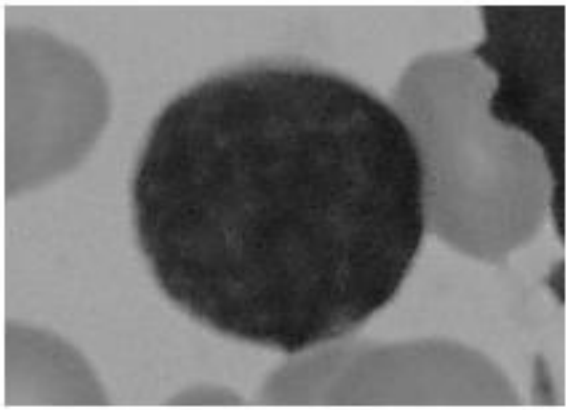

Figura 4-21: Linfocito aislado positivo para LLA

el objeto teniendo en cuenta que solo hay dos valores posibles (cero y uno), siendo uno blanco y, cero negro, matemáticamente se puede representar de la siguiente manera:

$$area = \sum_{n=1}^{N} p(n) \tag{4-5}$$

Donde p es el píxel, n su posición y, N es la cantidad total de los píxeles del área representativa del candidato, se realiza una suma total de todos los píxeles que conforman el objeto, es decir solo aquellos con valor uno.

- Excentricidad: los objetos que no son completamente uniformes pueden llegar a ser objetos que no sean linfocitos por lo que es necesario tenerlos en cuenta para removerlos de la clasificación.

  Este valor se calcula como la distancia entre los puntos focales de la elipse y la longitud del eje principal es decir:

$$Excentricidad = \frac{\sqrt{(Pf_1)^2 - (Pf_2)^2}}{LP} \tag{4-6}$$

  Donde Pf son los puntos focales de la elipse del objeto y LP es la longitud del eje principal el cual siempre será el de mayor longitud de la elipse, cuando estos dos valores son iguales y da como resultado uno, quiere decir que el objeto es una línea, si son completamente diferentes y el resultado se acerca a cero, es un objeto circular.

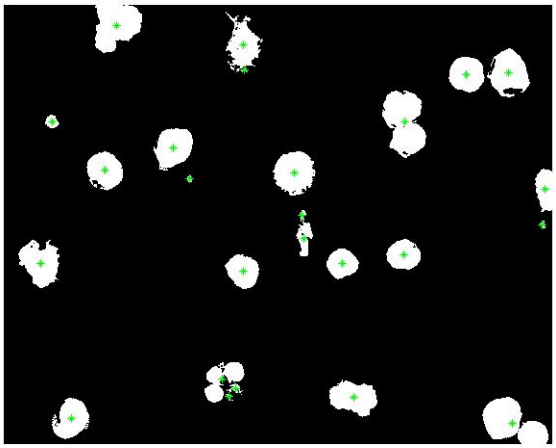

Figura 4-22: Centroides ubicados dentro de los objetos encontrados

- Solidez: como acercamiento a un análisis del citoplasma se tiene en cuenta la cantidad de píxeles que se encuentran cerca de los limites del objeto en proporción con la cantidad de píxeles del objeto.

  Matemáticamente, es la relación existente entre el área total y el área convexa del objeto, como se ve en la figura **4-24**, el área total es todo lo que es blanco en la imagen omitiendo el borde rojo y el área convexa es el área que esta dentro del borde rojo (es una representación, que no es como funciona, no es un área convexa real).

$$Solidez = \frac{A}{Ac} \tag{4-7}$$

  Donde A es el área total del objeto y Ac es la cantidad de píxeles que se encuentran dentro del área del casco convexo de la imagen.

- Área convexa: el área que representa el espacio entre la membrana y el núcleo, lo que permite realzar otros cálculos usando este valor como lo es la solidez

  Haciendo uso del casco convexo del objeto se calcula el área de esa figura interna. En el cual dentro de la figura **4-24** es el área que se encuentra dentro del la periferia roja.

4.6.1. Evaluación de características

Como ya se obtuvieron las características asociadas a los objetos candidatos, es importante determinar la capacidad de las mismas para discriminar entre las dos posibles clases. Para

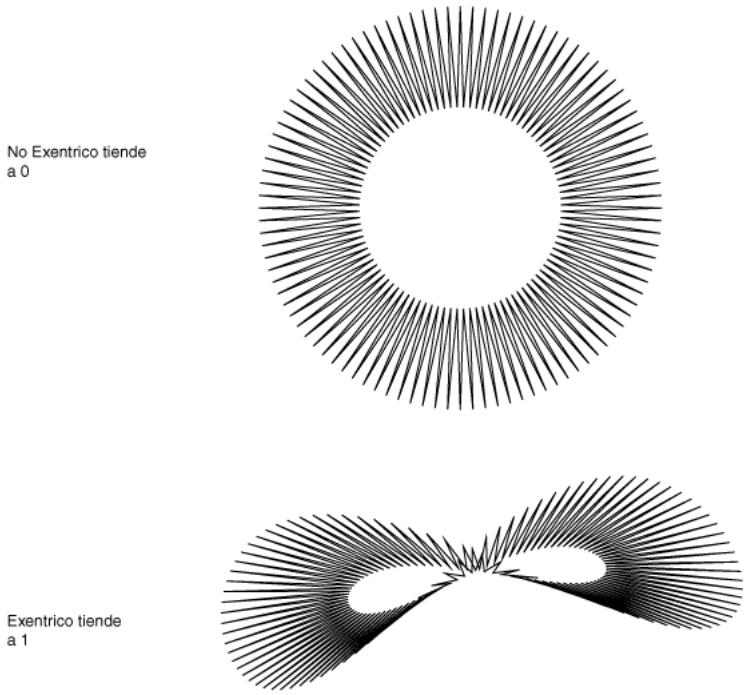

Figura 4-23: Representación de excentricidad, en la imagen superior excentricidad tiende a cero, mientras que, en la figura inferior tiende a uno

cumplir con este objetivo se realizó un estudio estadístico de *T-Student* el cual se calcula según la formula 4-8.

$$T = \frac{\overline{X_1} - \overline{X_2}}{S_{x1x2} \times \sqrt{\frac{2}{n}}} \tag{4-8}$$

Donde $\overline{X}$ es la media de la muestral tanto de la clase 1 (positivo para LLA) como de la clase 2 (negativo a LLA), $S_{x_1x_2}$ es la desviación estándar combinada y n es el tamaño de la muestra.

La idea detrás de este test es la evaluación de la posibilidad de que las características extraídas (que de entrada se suponen estar normalmente distribuidas) tiene funciones de distribución de probabilidad con medias diferentes y varianzas conocidas. La anterior afirmación se traduce en la verificación de la independencia de los valores de cada característica entre los dos posibles diagnósticos.

Para este ejercicio de comprobación se implementó un t-test(nombre alterno de test t-student)para variable simple, tomando como entrada la diferencia (resta) de los valores de una misma característica X, para las dos posibles clases de nuestro problema X_1 y X_2.

En este caso la decisión de prueba de hipótesis nula (h en nuestro caso) se concentra en

Figura 4-24: Representación del área total y el área convexa

la aprobación ($h = 0$) si es que la diferencia entre las distribuciones de probabilidad tiene media cero y describe una distribución normal, sin embargo, para nuestro problema es más importante el rechazo de esta hipótesis (hipótesis alternativa), pues al querer demostrar separabilidad en las muestras (características) resulta deseable que dichos vectores no tengan ni media cero (0) ni distribución normal, tal como se describe en la figura **4-25**.

Hipotesis Nula	H0: $\mu = \mu_0$	H0 $\mu \geq \mu_0$	H0: $\mu \leq \mu_0$
Hipotesis Alternativa	H1: $\mu \neq \mu_0$	H1: $\mu < \mu_0$	H1: $\mu > \mu_0$
Representacion grafica			

Figura 4-25: Representación de las posibles diferencias existentes entre gráficas gaussianas similares, donde se puede apreciar que el espacio en blanco pertenece a la hipótesis nula y el espacio sombreado a la hipótesis alterna

Los valores de las hipótesis tanto nula como alternativa se calculan haciendo uso de diferencia de la varianza que tienen los datos de entrada como se ve en la figura **4-25**, donde 0 significa que el sistema no rechaza la hipótesis nula. El objetivo de este análisis busca que se rechace la hipótesis nula indicando que estos tienen una separabilidad aceptable, donde entre mas bajo es el valor de T mucho mas lejos están o mas improbable es que se encuentren los valores de entrada.

Los valores T bajos ($< 0{,}05$) significan que la probabilidad de ocurrencia de que las características de positivos y negativos no signifiquen lo mismo para el sistema de clasificación; por lo que los valores mas cercanos a cero representan una baja probabilidad permitiendo

	Excentricidad	Área convexa	Área de relleno	Solidez
Estado (h)	0	1	0	1
valor T	0,1595	$5,88x10^{-6}$	0,4325	$3,4014x10^{-4}$

Tabla 4-2: Resultados para cada característica el T-Student

afirmar que son óptimos para su uso en clasificación.

Como se puede ver en la tabla **4-2** los resultados de la prueba T-student evidencian que las mejores características en términos de separabilidad estadística de las clases son *Área convexa y Solidez*. Esta afirmación hecha con el respaldo de este test también es confrmada a través de los boxplots expuestos en la figura **4-26**.

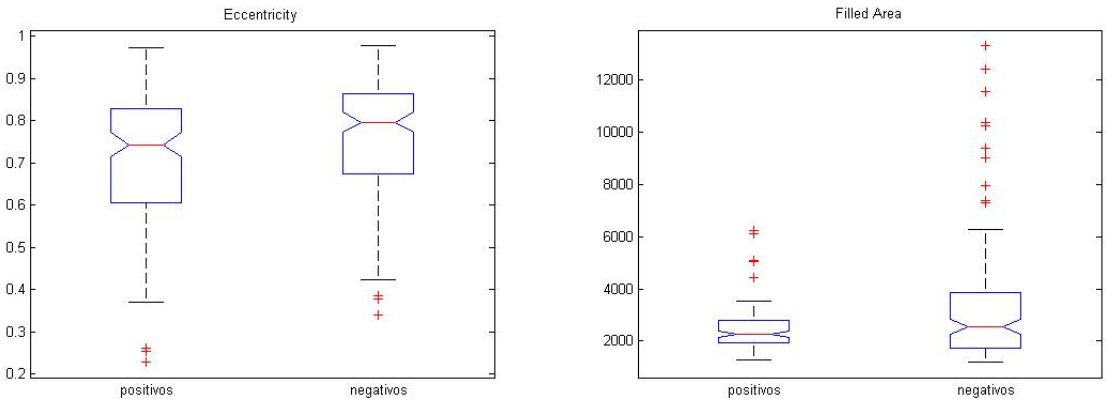

Figura 4-26: Boxplot, características no escogidas

Como se observa en los boxplots estos me indican como se distribuyen los datos entre positivos y negativos, los cuales me indican por porcentil la distribución de los datos y su desviación máxima. Para este caso se requiere que no se toquen los porcentil uno y tres los cuales representan la mayoría de los datos [IBM, 2018], debido a que si estos se llegases a tocar indicaría que los valores tanto positivos como negativos pueden pertenecer a ambos grupos en un punto dado, evitando una distinción en los datos, confirmando una hipótesis nula en el T-stundet.

Las características anteriormente mencionadas que en el T-student rechazaron la hipótesis nula (Área convexa y Solidez), son representativas para el profesional especializado, lo que indica que siempre tienen que ser verificadas por el profesional de curso. Es importante que estos descriptores resulten ser representaciones de los criterios presentados por la FAB, por lo que tiene importancia la validación final del especialista.

4.7. Clasificación

Las cantidades de datos de entrada en las características se escogieron de tal manera que existiera una cantidad similar de datos positivos y negativos, siendo estos 43 positivos y 42 negativos, lo que permitió realizar una base de datos de características amplia y variada entre positivas y negativas asegurando un balance entre las clases (sanos y no sanos).

En el caso de los datos de entrenamiento o de referencia, se usaron las etiquetas impuestas por los especialistas dentro de la ALL-DB1, donde se separaron todos aquellos linfocitos pertenecientes a imágenes diagnosticadas como "no sanas"de los pertenecientes a las imágenes "sanas"; con estos dos conjuntos se crearon dos grupos de características de linfocitos de ambas clases.

Se decidió usar una maquina de soporte vectorial en razón a su capacidad de trabajo óptimo bajo condiciones de características no lineales, como lo son las señales de imagen que se procesaron durante el desarrollo del trabajo de grado; además, de permitir trabajar rápidamente con entradas poco convencionales y evitar el sobre ajuste, lo que permitió acercarse al método manual ejercido por el profesional.

Este método evalúa las características previamente seleccionadas, haciendo uso de un hiper plano que permite separarlas y luego clasificarlas; estos hiper planos se calculan maximizando la distancia entre las características y cada uno de ellos, evaluándolos todos y seleccionado el mejor, como se evidencia en la figura **4-27**.

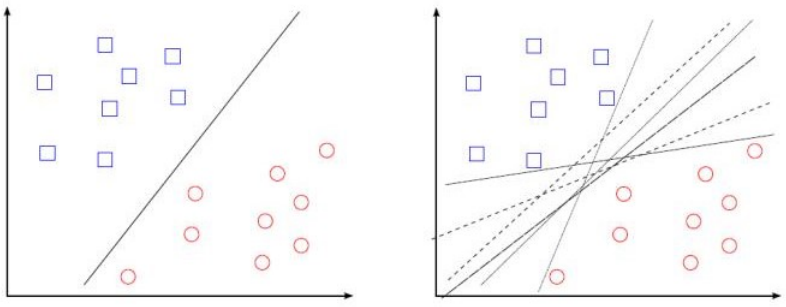

Figura 4-27: Ilustración de posibles planos para separar dos características

De todos los planos el sistema elije el mejor según los parámetros dados y ajustados usualmente por una función matemática llamada kernel, que permite configurar la flexibilidad del sistema, ya sea permitiendo en algunos casos datos mal clasificados para evitar sobre ajuste o, haciendo el sistema mas rígido al escoger el hiper plano, todo esto depende de las

características y su clasificación de entrada.

La máquina de soporte vectorial es una técnica de clasificación, basada en el aprendizaje estadístico y con el mínimo riesgo estructural, la cual es aplicada principalmente para tareas de clasificación binaria [Carreño, 2014], el uso de funciones kernel y un rendimiento matemático efectivo, permiten a las máquinas de soporte vectorial crear hiper planos que a su vez aumentan la distancia entre las clases a clasificar, al mismo tiempo que disminuyen el error de clasificación total [Zararsiz, 2012].

$$f(x) = x\beta + \beta_0 x = (0, 1, 2, 3, 4, ...N - 1) \tag{4-9}$$

Donde x es el valor de entrada, y β y β_0 son parámetros de sintonización de la máquina de soporte vectorial.

La anterior función representa cómo es una máquina de soporte vectorial matemáticamente, teniendo en cuenta que el beta es el tipo de clasificador usado, que en este caso es un clasificador log-lig con la siguiente fórmula matemática:

$$\frac{1}{1 + \epsilon^{-}x} = \beta \tag{4-10}$$

También se realizó una validación cruzada entre las entradas al sistema, para garantizar que las entradas de entrenamiento y validación sean independientes, con una partición cada cinco (5) datos, evaluando cada grupo por separado y asegurando una buena distribución en los datos.

De esa manera se logra, por medio de las características de aprendizaje y las ecuaciones anteriores, un modelo de predicción para los datos siguientes y adquirir un sistema automático de clasificación binaria.

Se usó este sistema de clasificación, con diferentes kernels: RBF, lineal, polinomial y gaussiano, esto para evaluar cual de estos anteriores presenta las mejores métricas de evaluación.

Los kernel anteriores son modificadores del clasificador debido a que cada una representa métodos diferentes para la generación de hiper planos debido a que representan diferentes ecuaciones con diferentes resultados, estas ecuaciones son:

Para el kernel RBF y gaussiano se representa de la siguiente forma, donde x_1 y x_2 son los valores de las clases de entrada y α es la anchura del kernel

$$K(x_1, x_2) = \exp -\frac{||x_1 - x_2||^2}{2\alpha^2} \tag{4-11}$$

Para el kernel lineal

$$K(x_1, x_2) = x_1^T x_2 \tag{4-12}$$

Para el kernel polinómico donde ρ es el tamaño mínimo del polinomio

$$K(x_1, x_2) = (x_1^T x_2 + 1)^\rho \tag{4-13}$$

Cada uno de estos modifica el modo de aprendizaje del la maquina de soporte vectorial permitiendo mas o menos errores de la clasificación para adquirir mejores vectores de separación de clases cambiando el resultado final a menos que las características de entrada sean lo suficientemente separadas evitando diferencias significativas en el resultado final

4.8. Metodología de evaluación

Para realizar una validación de los datos entregados por el modelo de predicción se dividieron los datos en porcentajes de 70-30, donde el 70 % se usó para entrenar el sistema y para la validación cruzada en la sección de clasificación. El 30 % de los datos se empleó como ciego estadístico, este se usó para comprobar el funcionamiento del sistema.

El sistema fue evaluado bajo las siguientes métricas: tabla de confusión para visualizar la viabilidad del sistema, como se observa en la tabla **4-3**, además, se realizaran pruebas con diferentes resultados del modelos de predicción variando el kernel del método de SVM.

Verdaderos positivos	Verdaderos negativos
Falsos Positivos	Falsos Negativos

Tabla 4-3: Tabla de confusión

Dentro de la tabla de confusión, los valores adquiridos permitieron calcular los indicadores requeridos (sensibilidad, especificidad, precisión y exactitud) y con ellos, se tiene el criterio necesario para evaluar objetivamente el resultado de la predicción.

Los valores se calcularon teniendo en cuenta la siguiente nomenclatura; Vp son los datos enfermos y el modelo lo predijo correctamente; Vn son aquellos datos sanos y el modelo acertó al decir que son negativos; Fp son los valores que a pesar de ser sanos el modelo de

predicción los clasificó como enfermos y, Fn son los candidatos que aunque estén enfermos el sistema no los determinó de esa manera, las fórmulas utilizadas para adquirir las métricas fueron las siguientes:

$$Sensibilidad = \frac{Vp}{Vp + Fn} \tag{4-14}$$

Sensibilidad se define como la probabilidad que tiene un sistema de poder clasificar las entradas con una etiqueta de positivo para LLA de manera correcta, esto quiere decir que en caso de que no existan valores de falsos negativos en el sistema, evaluará correctamente todos los casos donde se tenga la enfermedad.

$$Especificidad = \frac{Vn}{Vn + Fp} \tag{4-15}$$

Especificidad se define como la probabilidad de organizar correctamente los valores de entrada que sean clasificados como negativos para LLA, esta función es inversa a la sensibilidad, dando la capacidad de detectar los valores negativos correctamente, este valor depende de la cantidad de verdaderos negativos que se tenga de entrada y de los falsos positivos.

$$Precisión = \frac{VP}{Vp + Fp} \tag{4-16}$$

Precisión es la probabilidad de acertar correctamente los valores de LLA positivos, teniendo en cuenta cuantos de esos clasificó de manera errónea, dando como resultado la cantidad de valores reales definidos como tales.

$$Exactitud = \frac{Vp + Vn}{Vp + Vn + Fp + Fn} \tag{4-17}$$

Exactitud es la probabilidad que tiene el sistema de acertar en general, teniendo en cuenta cuanto clasificó correctamente y cuanto no, esto quiere decir que en caso de que exista un sistema sin margen de error este valor sera uno (1).

Al realizar un estudio por medio de una máquina de soporte vectorial se hizo una comparación de diferentes ámbitos de aprendizaje para demostrar la diferencia entre las características seleccionadas y los diferentes kernel que podrían afectar el método, como se visualiza en las tablas de evaluación de diferentes kernel **4-5** y **4-7** con sus correspondientes métricas presentadas en las tablas de confusión **4-4** y **4-6** respectivamente.

Como se evidencia en la tabla anterior **4-5** las métricas dadas por las características de excentricidad y área de relleno están por debajo de lo requerido, ya que se tiene como referente

Predicción	RBF	Lineal	polynomial	Gaussiano
Vp	24	17	24	26
Vn	43	43	43	43
Fp	19	27	19	17
Fn	0	0	0	0

Tabla 4-4: Tablas de confusión con las características de Excentricidad y Área de relleno, las menos significativas segun el T-student

Métrica	RBF	Lineal	Polinomial	Gaussiano
Exactitud	77.9069	69,7674	77,9069	80,2325
Sensibilidad	100	100	100	100
Especificidad	69,3548	62,3188	69,3548	71.6667
Precisión	55,8139	39,5348	55,81	60,4651

Tabla 4-5: Evaluación de diferentes kernel de las características Excentricidad y Área de relleno, las menos significativas segun el T-student

al profesional que tiene una precisión entre el 60 % y 70 % [Amin M, 2015] y estar por debajo de estos estándares no es óptimo para el sistema que se quiere realizar.

Se observa que la sensibilidad en la tablas **4-5** y **4-7** siempre es del 100 %, lo que da a entender que siempre clasificará correctamente los candidatos enfermos del sistema.

Contrario a la sensibilidad, la especificidad en la tabla **4-5** muestra que la capacidad que tienen estas características (excentricidad y área de relleno) de clasificar candidatos positivos se encuentra entre un 69 % y un 71 %, indicando que solo este porcentaje de los candidatos reales sanos fue clasificado correctamente y que aun siendo superior al 65 %, sigue bajo para los objetivos del proyecto.

En la tabla **4-7** se observa que las características de área convexa y solidez están en va-

Predicción	RBF	Lineal	polynomial	Gaussiano
Vp	34	34	34	33
Vn	43	43	43	43
Fp	9	9	9	10
Fn	0	0	0	0

Tabla 4-6: Tablas de confusión con las las características de Área convexa y Solidez, las mas significativas según el T-student

Métrica	RBF	Lineal	Polinomial	Gaussiano
Exactitud	89,5348	89,5348	89,5348	88,3721
Sensibilidad	100	100	100	100
Especificidad	82,6923	82,6923	82,6923	81,1321
Precisión	79,0697	79,0697	79,0697	76,7441

Tabla 4-7: Evaluación de diferentes kernel con las las características de Área convexa y Solidez, las mas significativas según el T-student

lores entre el 81 % y el 82 %, indicando que la mayoría de los datos de entrada sanos fueron correctamente clasificados, con un margen de error cercano al 20 %.

Observando los valores de precisión de ambas tablas, se tiene que los valores de la tabla **4-7** son ampliamente favorables debido a que están cerca al 80 % de precisión, respecto de los resultados obtenidos en la tabla **4-5**, que en ningún caso superaron el 60 %, siendo una probabilidad casi al azar de acertar en la clasificación.
Para los valores de la tabla **4-7**, los indicadores de sensibilidad y especificidad son altos,lo que indica que la capacidad que tienen estas características para clasificar correctamente es alta, lo cual se refleja en su exactitud y su precisión.

Observando los resultados obtenidos con los diferentes kernel en el SVM, a pesar que se tuvieron resultados similares, el tiempo de ejecución de cada uno fue diferente, siendo el RBF el mas veloz y el polinomial el mas lento, por lo que se deduce que el mejor en este caso es el RBF.

Se realizo una prueba de sintonización con 43 entradas positivas para LLA y 43 entradas negativas LLA, donde se variaban los valores de peso de α de un 10 % , 30 % y 50 % del los valores de entrada, con diferentes kernel para comprobar su eficacia, dando como resultado la tabla **4-8**

Se realizó una evaluación con todas las características posibles pero el sistema no podía evalúa entregando siempre en todos los kernel daba una clasificación errónea, similar a lo sucedió en la tabla **4-8** donde no lograba entregar resultados congruentes ni útiles para una interpretación.

Por último, se realizó una curva ROC (Reciber Operating Characteristics - ROC, por sus siglas en inglés) la cual es una prueba de diagnóstico para modelos de clasificación a través de la especificidad y sensibilidad del mismo; para el proyecto de grado se obtuvo el el resultado que se muestra en la figura **4-28**.

10 %	RBF	lineal	polynomial	gaussiano
Vp	0	0	0	0
Vn	43	43	43	43
Fp	0	0	0	0
Fn	43	43	43	43
30 %	RBF	lineal	polynomial	gaussiano
Vp	0	0	0	0
Vn	43	43	43	43
Fp	0	0	0	0
Fn	43	43	43	43
Vp	0	0	0	0
50 %	RBF	lineal	polynomial	gaussiano
Vn	43	43	43	43
Fp	0	0	0	0
Fn	43	43	43	43

Tabla 4-8: sintonización con valores del porcentuales de los valores de entrada

Los valores finales del modelo de clasificación adquirieron un 82 % de área bajo la curva, lo que indica que si el sistema se repite bajo las mismas condiciones a un sujeto enfermo o positivo para LLA, el 82 % de las veces será catalogado como enfermo; de esta prueba no se tiene un margen específico para definir el rendimiento de un sistema, pero se reconoce que entre mas cercano a 100 % es mejor [Jaime Cerda, 2012].

La curva ROC anterior también indica que hay un 18 % de probabilidad de tener un falso positivo o un falso negativo en el sistema, el cual es un rango aceptable en comparación con el error del especialista el cual es de un 30 % al 40 %.

Los resultados de los métodos expuestos anteriormente, tablas de confusión y curva ROC, indican un desarrollo exitoso del proyecto al tener métricas por encima de los reportados por los profesionales.

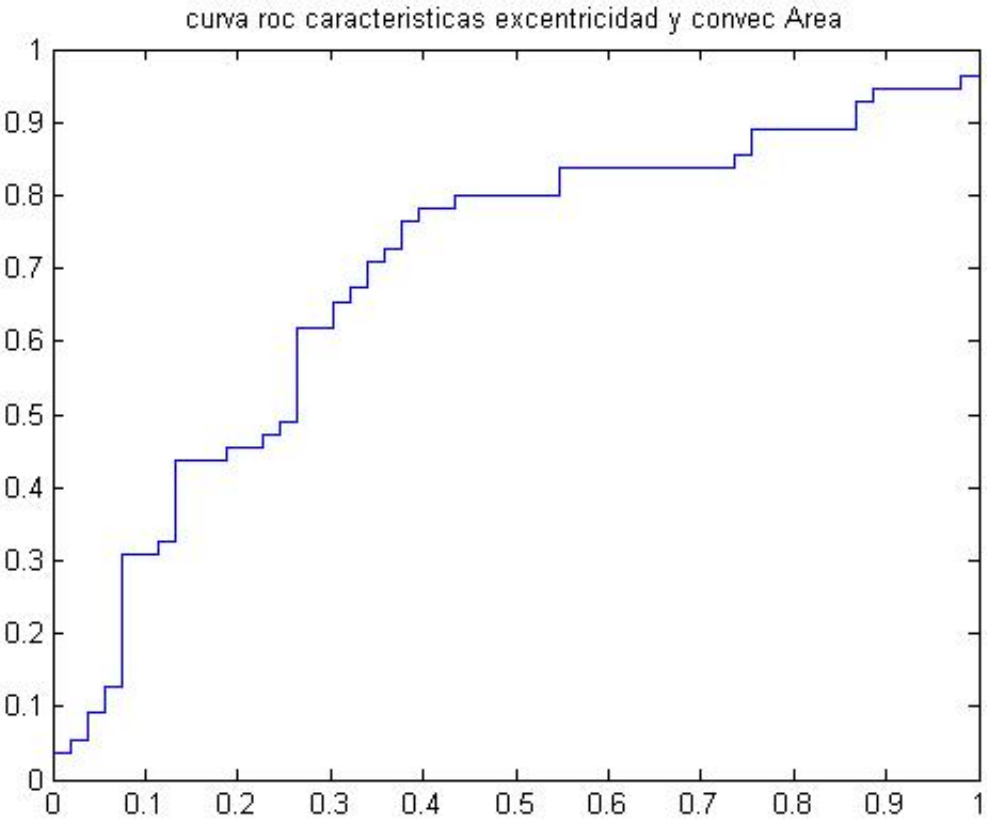

Figura 4-28: Curva ROC, areá bajo la curva igual a 82 %

5 Conclusiones y recomendaciones

A lo largo del desarrollo del trabajo de grado y el análisis de resultados se dieron diferentes tipos de conclusiones que son las que serán descritas a continuación:

- Al realizar un método que emplea características usadas por el especialista, da interpretabilidad de los resultados para el especialista sin llegar a cegarlo en el diagnostico.

- Los resultados obtenidos cumplen satisfactoriamente la tarea de clasificar linfocitos con un índice mayor a la del profesional, con una metodología similar. Este índice pasa de ser de un rango de 30 % - 40 % de error, al 20 % - 21 %.

- Una vez se realizaron las evaluaciones con diferentes kernel del SVM se concluyó que las características dadas por el T-student, son discriminantes bajo todos los kernel, debido a su similitud en los indicadores de la matriz de confusión.

- Dado lo anterior, el objetivo de la tesis se cumplió al realizar un método automático de clasificación de linfocitos afectados por leucemia linfoblástica aguda en imágenes hematológicas a través de un algoritmo.

- El modelo trabajado funciona al servir como herramienta de apoyo al profesional; debe ser desarrollado a profundidad y complementado con una metodología de adquisición de imágenes.

- A pesar que requiere estudio clínicos para poder tener un funcionamiento completo para determinar su eficacia en campo, puede ser usado de base para otros trabajos de grado.

Bibliografía

[Amin M, 2015] Amin M, Kermani S, T. A. O. M. (2015). Recognition of acute lymphoblastic leukemia cells in microscopic images using k-means clustering and support vector machine classifier. *Journal of Medical Signals and Sensors*, 5:49–58.

[Ariffin, 2012] Ariffin (2012). An image processng aplications for the localization and segmentacion of lymphoblas cells using peropheral bloob images. *springer*.

[Biau, 2012] Biau, G. (2012). Analysis of a random forests model. *Journal of Machine Learning Research*, 13:1063–1095.

[Brummel, 2002] Brummel (2002). Thrombin functions during tissue factor–induced blood coagulation. *HEMOSTASIS, THROMBOSIS, AND VASCULAR BIOLOGY*, 100:148–152.

[Campos, 2017] Campos, V. P. G. (2017). *Modelado medieante Random forests de las emisiones de autobuses urbanos*. Universidad Politecnica de Madrid, primera edicion edition.

[Carreño, 2014] Carreño (2014). Errores en la formulación de quimioteriapia. *Revista colombiana de cancerologia*, 1(44):179,185.

[del pilar, 2016] del pilar, C. (2016). caracterización clínico epidemiológica de los pacientes pediatricos con leicemias agudas en la clínica universitaria colombia serie de casos 2011-2014. *revista de pediatria de colombia*, 1(2):17,22.

[Departamento de Ingeniería Electrónica, 2005] Departamento de Ingeniería Electrónica, T. y. A. (2005).

[Dr.Karthikeyan, 2017] Dr.Karthikeyan (2017). Microscopic image segmentacion using fuzzy c-means for leukemia diagnosis. *internacional journal of advance research in science,engineering and tecnology*.

[E.Cuevas, 2010] E.Cuevas (2010). Segmentacion y deteccion de globulos blancos en imagenes usando sistemas inmunes artificiales. *Revista mexicana de ingenieria biomedica*, 2:119–134.

[Enrique J, 2014] Enrique J, C. S. (2014). Tutorial sobre maquinas de vectores soporte (svm). *Universidad Nacional de edicacion a distancia ()UNEd*.

[Gersten, 2018] Gersten, T. (2018). centros oncologicos para niños.

[González, 2008] González, R. C. (2008). *Digital image Processing.* Pearson.

[H, 2012] H, M. (2012). An image processing application for the localization and segmentation of lymphoblast cell using peripheral blood images. $Med_Syst, 36 : 2149 - -2158$.

[H, 2017] H, M. (2017). Disease diagnosis using rbcs & wbcs cell structure by image processing. *International Journal of Scientific Research in Science and Technology.*, 3(2):120–123.

[Hamid, 2013] Hamid, G. A. (2013). *CLINICAL HEAMTOLOGY.* Eden university.

[IBM, 2018] IBM (2018).

[Jagadeesh, 2013] Jagadeesh, S. (2013). Image processing bases approach to cancel cell prediction on blood samples. *Internacional journal of technology and engineering sciences.*

[Jaime Cerda, 2012] Jaime Cerda, L. C. (2012). Uso de curvas roc en investigación clínica. aspectos teórico-prácticos. *Revista chilena de infectología*, 2:138–141.

[Jairo Aguilera López, 2015] Jairo Aguilera López, E. a. (2015). *Anàlisis de situación del cáncer en Colombia 2015.* Insituto Colombiano de Cancerología.

[Jiménez, 2008] Jiménez, M. (2008). introducción al tratamiento digital y clustering de imágenes. *REE.*

[Kandil, 2016] Kandil, A. (2016). Automatic segmentation of acute leukemia cells. *International Journal of Computer Applications*, 133(10).

[Labaty, 2011] Labaty, r. (2011). All-ibd the acute lymphoblastic leukemia image database for image processing. *IEEE*, 65.

[Lillo, 2012] Lillo, S. (2012). La sangre.

[Mathworks, 2016] Mathworks (2016). Global image treshold.

[Miguel A. Castillo Martínez, 2016] Miguel A. Castillo Martínez, e. a. (2016). Preprocesamiento de imágenes dermatoscopicas para extracción de características. *Research in Computing Science.*

[Miralles, 2017] Miralles, S. R. (2017). *Análisis de imágenes digitales de células linfoides de sangre periférica a partir de microscopia óptica.* Universidad Politécnica de Cataluña.

[Mishra, 2017] Mishra, e. a. (2017). Gray level co-occurrence matrix and random forest based acute lymphoblastic leukemia detection. *Biomedical Signal Processing and Control*, 33:272–280.

[Morales, 2017] Morales, N. F. B. (2017). *Teledetección espacial*. Universidad Nacional Agraria de la Selva.

[Msalgobar, 2017] Msalgobar (2017). ¿que es la sangre.

[Ofarrin, 2014] Ofarrin (2014). Leucemia mieloide crónica en un adulto con inmunodeficiencia común variable. *Revista médica institucional mexicana*, 94:94–97.

[Piuri, 2004] Piuri, V. (2004). Morphological classificaction of blood leucocytes by microscope images. *Computationvl inteligence far memrement systems and applications*.

[PMfarma, 2015] PMfarma (2015). Leucemia linfoblastica aguda, tercer tipo de cancer hamtologico mas frecuente.

[Putzu L, 2014] Putzu L, Caocci, D. u. C. (2014). Leucocyte classification for leukaemia detection using image processing techniques. *Artificial Intelligence in Medicine*, 62:179–191.

[R., 2013] R., D. (2013). Infliximab inhibits activation and effector functions of peripheral blood t cells in vitro from patients with clinically active ulcerative colitis. *Human Immunology,*, 78(3):275–284.

[Sala, 2003] Sala, M. (2003). *Farmacia hospitalaria*, volume 2. Sociedad española de farmacia hospitalaria.

[Sarrafzadeh, 2015] Sarrafzadeh, O. (2015). Detecting different sub-types of acute mielogenous leukemia using dictionary learnig and sparce representation. *Internaconal conferente of image procesing*.

[Scotii, 2004] Scotii (2004). Morphological clasification of blood leuceocytes by microscope images. *IEEE international conference on computationvl inteligence far memrement systems and applications*, 1:14,16. CTMSA 2014.

[Scotti, 2006] Scotti (2006). Robust segmentacion and measurements techniques of white cells in blood microscope images. *Instrumentation and measurement*, 1(1):24–27. IMTC 2006.

[Scottii, 2005] Scottii (2005). Automatic morphological analysis for acute leukemia identification in peripher blood microscope images. *Computacional inteligence for measument systems and aplications*.

[Society, 2016] Society, A. C. (2016). *Detección temprana, diagnóstico tipos de leucemia*, chapter 3. American Cancer Society.

[S.Ordaz, 2011] S.Ordaz (2011). Deteccion de leucemia linfoblÁstica aguda usando lÓgica difusa y redes neuronales. *Instituto politecnico nacional*.

[Srisukkhama, 2017] Srisukkhama, W. (2017). Intelligent leukaemia diagnosis with bare-bones pso based feature optimization. *Applied Soft Computing*, 56:405–419.

[Suca, 2016] Suca, C. (2016). Comparación de algoritmos de clasificación para la predicción de casos de obesidad infantil. *Research Gate*.

[Vaguela, 2015] Vaguela (2015). Leukemia detection using digital image processing techniques. *International Journal of Applied Information Systems*, 10:43–51.

[Zararsiz, 2012] Zararsiz, G. (2012). Bagging support vector machines for leukemia classification. *IJCSI International Journal of Computer Science Issues*.