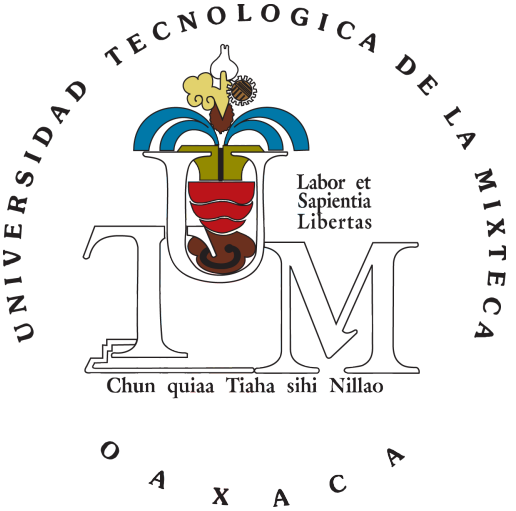


UNIVERSIDAD TECNOLÓGICA DE LA MIXTECA

RECONOCIMIENTO DE PATRONES DE LEUCOCITOS Y ERITROCITOS: IDENTIFICACIÓN, CONTEO Y CLASIFICACIÓN

TESIS

PARA OBTENER EL TÍTULO DE

INGENIERO EN COMPUTACIÓN

PRESENTA:

ODALYS PAZ MENDOZA

DIRECTOR DE TESIS:

DR. RAÚL CRUZ BARBOSA

ASESOR DE TESIS:

DR. ANTONIO ORANTES MOLINA

HUAJUAPAN DE LEÓN, OAXACA, SEPTIEMBRE DE 2019.

Dedicado a mis papás, Quirino y Julia,
por todo el apoyo que me han dado.

Dedicado a mis hermanos,
Miriam, Mauricio,
Jose, Alex y Jazmín.

†

Agradecimientos

A mis padres, por darme la gran oportunidad de estudiar una ingeniería. Siempre estaré muy agradecida con mi padre, que a pesar de la distancia me ha brindado su apoyo y la fortaleza necesaria para superarme cada día. A mi madre por sus enseñanzas y amor. Por todo esto y muchas cosas más, gracias.

Mi más sincero agradecimiento a mi director de tesis, el Dr. Raúl Cruz Barbosa, por motivarme a realizar este proyecto. Gracias por su tiempo, orientación y paciencia a lo largo de la realización de esta tesis. Toda mi admiración y respeto hacia usted.

Un agradecimiento especial a MTCA. Saiveth Hernández Hernández por sus consejos y su apoyo incondicional a lo largo de la realización de esta tesis.

A los especialistas, QFB. Elizabeth Rivera Galindo, EHDL. Jacob Santiago Blanco y QFB. Glafira Torres Gómez por ayudarme con lo relacionado en el área de hematología. Gracias por su ayuda desinteresada.

A mis amigas Alma, Lizbeth y Nancy por estar siempre presentes y por la complicidad a lo largo de estos años. A los compañeros que se formaron conmigo, gracias por su amistad y los buenos momentos juntos.

A mis familiares y amigos que me han brindado su amistad, consejos, ánimo y compañía en los momentos más difíciles. Arigatou!.

A mi asesor de tesis el Dr. Antonio Orantes, a mis sinodales, Dr. Eduardo Sánchez Soto, MTCA. Moisés E. Ramírez Guzmán y Dr. Rosebet Miranda, gracias por los comentarios y observaciones realizadas para la mejora de este trabajo.

A la Universidad Tecnológica de la Mixteca por darme la oportunidad de realizar mis estudios.

Índice general

Resumen **XIII**

1. Introducción **1**
- 1.1. Planteamiento del Problema . . . 3
- 1.2. Justificación . . . 5
- 1.3. Hipótesis . . . 6
- 1.4. Objetivos . . . 6
  - 1.4.1. Objetivo general . . . 6
  - 1.4.2. Objetivos específicos . . . 6
- 1.5. Metas . . . 7
- 1.6. Trabajos relacionados . . . 7
- 1.7. Metodología . . . 9

2. Marco Teórico **13**
- 2.1. Frotis y células sanguíneas . . . 13
- 2.2. Procesamiento Digital de Imágenes . . . 14
  - 2.2.1. Etapas del Procesamiento Digital de Imágenes . . . 15
  - 2.2.2. Preprocesamiento de imágenes digitales de frotis sanguíneo . . . 16
  - 2.2.3. Técnicas utilizadas para identificar leucocitos y eritrocitos . . . 17
  - 2.2.4. Técnicas utilizadas para contar leucocitos y eritrocitos 22
- 2.3. Reconocimiento de Patrones . . . 22
  - 2.3.1. Componentes de un Sistema de Reconocimiento de Patrones . . . 23
  - 2.3.2. Extracción de características . . . 24
  - 2.3.3. Reconocimiento de leucocitos . . . 28
- 2.4. Enfoques de clasificación de leucocitos . . . 32

3. Desarrollo del proyecto **33**
3.1. Especificaciones de hardware y software 33
3.2. Módulos del proyecto 34
3.2.1. Identificación y conteo de leucocitos y eritrocitos 34
3.2.2. Reconocimiento de leucocitos 43
3.3. Aplicación de muestra 50

4. Resultados **55**
4.1. Conjunto de datos 55
4.2. Resultados de identificación y conteo 57
4.2.1. Identificación de núcleos de leucocitos 59
4.2.2. Identificación de células sanguíneas 62
4.2.3. Identificación y conteo de leucocitos y eritrocitos 67
4.3. Reconocimiento de leucocitos 73
4.3.1. Enfoques de clasificación de leucocitos 74
4.3.2. Extracción de características 75
4.3.3. Resultados de reconocimiento de leucocitos 76

5. Conclusiones y trabajo a futuro **89**

Bibliografía **90**

Anexos **98**

A. Células sanguíneas **99**
A.1. Frotis sanguíneo 99
A.2. Leucocitos 100
A.3. Eritrocitos 103

B. Características seleccionadas para el reconocimiento de leucocitos **105**

C. Configuración y manual de usuario de la aplicación Web **111**
C.1. Configuración 111
C.2. Manual de usuario 114

Índice de figuras

1.1. Metodología para la identificación, conteo de leucocitos y eritrocitos, asimismo para el reconocimiento de 5 tipos de leucocitos. 11

2.1. Etapas del Procesamiento Digital de Imágenes. 15
2.2. Minimos Locales, Cuencas y Lineas de Watershed. 20
2.3. Componentes de un Sistema de Reconocimiento de Patrones. . 24

3.1. Imagenen ilustrativa de un frotis sanguineo y su correspondiente segmentación de núcleos. a) Frotis sanguíneo. b) Núcleos segmentados. 35
3.2. Diagrama del proceso para la identificación de núcleos. 35
3.3. Imágenes ilustrativas a) frotis sanguíneo, b) regiones fusionadas, c) segmentación ideal. 37
3.4. Diagrama del proceso para la identificación de células sanguíneas. 37
3.5. Diagrama del proceso para la identificación de leucocitos y eritrocitos. 39
3.6. Proceso de creación de la máscara de segmentación conjunta. . 40
3.7. Diagrama del proceso para el conteo de leucocitos y eritrocitos. 41
3.8. Candidatos a leucocitos identificados en un rectángulo. 42
3.9. Imágenes de leucocitos aislados (primera fila) y máscaras de segmentación (segunda fila). 42
3.10. Diagrama del proceso para la generación de características. . . 44
3.11. Imágenes de las regiones de interés del leucocito. 44
3.12. Diagrama proceso de la selección de características. a) Genera sólo un subconjunto de k características. b) Genera múltiples subconjuntos de k características. 47

3.13. Diagrama proceso para la evaluación de cada subconjunto de características. 48
3.14. Diagrama del proceso para la generación de modelos. 49
3.15. Arquitectura de la aplicación desarrollada. 51
3.16. Diagrama del proceso para el reconocimiento de leucocitos. . . 53

4.1. Imagen de frotis sanguíneo de calidad a) alta, b) media y c) baja calidad. 58
4.2. Ejemplo del proceso del preprocesamiento y segmentación. a) Imagen de alta calidad. b) Filtro mediana. b) Conversión a escala de grises. d) Umbralización de Yen. 59
4.3. Resultados de la identificación de núcleos en imágenes, en orden descendente por renglón: alta, media y baja calidad. a) Núcleos segmentados. b) Núcleos superpuestos a la imagen original. 61
4.4. Ejemplo del preprocesamiento para la imagen de alta calidad. a) Filtro mediana. b) Conversión a escala de grises. c) *CLAHE*. d) Transformación gamma. 62
4.5. Resultados de segmentación de células en imágenes, en orden descendente por renglón: alta, media y baja calidad. a) Umbralización global de Otsu. b) I_{seg}. 64
4.6. Resultados de postprocesamiento para separar regiones en imágenes, en orden descendente por renglón: alta, media y baja calidad. a) I_{seg} con etiquetas. b) Transformación de *watershed*. 66
4.7. Resultados de identificación de eritrocitos y leucocitos en imágenes, en orden descendente por renglón: alta, media y baja calidad. a) Leucocitos. b) Eritrocitos. 68
4.8. Máscara de segmentación conjunta para la imagen de a) alta, b) media y c) baja calidad. 69
4.9. Resultados de conteo de leucocitos y eritrocitos en imágenes de a) alta, b) media y c) baja calidad. 70
4.10. Imágenes de leucocitos aislados (primera fila) y máscaras de segmentación (segunda fila). 71
4.11. Neutrófilo con baja calidad de tinción en el citoplasma. 72
4.12. Enfoques de clasificación de leucocitos. Clasificación mediante a) cinco clases principales y b) por tipos y subtipos de leucocitos. 74

4.13. Generación de datos de entrenamiento y prueba: conjunto de datos para el enfoque con 5 clases, tipo (3 clases) y subtipo (3 clases) a) DS_{5C}, b) DS_{Tipo} y c) $DS_{subtipo}$. El símbolo C1' significa la agrupación de tres clases en el tipo granulocito. . . 78

A.1. Frotis de sangre periférica teñido de manera óptima que demuestra la zona adecuada para realizar la fórmula diferencial de leucocitos. 99
A.2. Neutrófilo a) banda y b) segmentado. 101
A.3. Basófilos. 102
A.4. Eosinófilos. 102
A.5. Linfocitos. 103
A.6. Monocito con a) vacuolas y b) sin vacuolas. 103

C.1. Ejemplo de cómo iniciar la API-REST. 113
C.2. Ejemplo de cómo iniciar la aplicación Web. 113
C.3. Ingresar a la aplicación desde el navegador Web. 114
C.4. Página principal de la aplicación. 114
C.5. Página para subir la imagen de frotis sanguíneo. 115
C.6. Resultados: a) conteo de eritrocitos y leucocitos, y b) leucocitos identificados. 116
C.7. Página de espera. 117
C.8. Página de resultados de leucocitos. 117

Índice de cuadros

3.1. Características del Equipo. 33
3.2. Funciones principales del módulo identificación de núcleos. . . 36
3.3. Funciones principales del módulo identificación de células sanguíneas. 38
3.4. Módulo identificación de leucocitos y eritrocitos. 41
3.5. Funciones principales del módulo de conteo. 43
3.6. Lista de características utilizadas para generar el vector de características. 45
3.7. Funciones principales del módulo de selección de características. 48
3.8. Funciones principales del módulo de generación de modelos. . 50
3.9. Rutas más importantes para la navegación en la interfaz. . . . 52
3.10. End-point más importantes de la API-REST. 53

4.1. Distribución de leucocitos del conjunto DS_1 respecto a una clasificación de cinco clases. 56
4.2. Distribución de leucocitos del conjunto DS_2 respecto a una clasificación de cinco clases. 56
4.3. Distribución de leucocitos de la unión de DS_1 y DS_2. 57
4.4. Distribución de leucocitos para $Union_A$. 57
4.5. Conteo de regiones candidatas a leucocitos en imágenes: comparativa de conteo manual vs conteo automático. 71
4.6. Conteo de eritrocitos: comparativa de conteo manual vs conteo automático. 73
4.7. Distribución de ejemplos por conjuntos de imágenes. 75
4.8. Distribución de leucocitos, respecto al enfoque de clasificación con cinco clases. 76
4.9. Distribución de leucocitos, respecto al enfoque de clasificación en dos etapas: tipos. 76

4.10. Distribución de leucocitos, respecto al enfoque de clasificación en dos etapas: subtipos. 76
4.11. Parámetros de los algoritmos de clasificación. 79
4.12. Resultados de medidas de evaluación de los modelos para el reconocimiento de leucocitos para el caso cuando la distribución de clases es en basófilo, eosinófilo, linfocito, monocito y neutrófilo. 80
4.13. Resultados de medidas de evaluación de los modelos para el reconocimiento de leucocitos para el caso cuando la distribución de clases es por tipo, es decir, linfocito, monocito y granulocito. 80
4.14. Resultados de medidas de evaluación de los modelos para el reconocimiento de leucocitos para el caso cuando la distribución de clases es por subtipo, es decir, basófilo, eosinófilo, y neutrófilo. 80
4.15. Resultados de selección de características y medidas de evaluación de los modelos para el reconocimiento de leucocitos para cinco clases. 83
4.16. Resultados de selección de características y medidas de evaluación de los modelos para el reconocimiento de leucocitos para tres clases: tipo. 84
4.17. Resultados de selección de características y medidas de evaluación de los modelos para el reconocimiento de leucocitos para tres clases: subtipo. 84
4.18. Resumen de los mejores modelos para los tres tipos de enfoque de clasificación. 85
4.19. Matriz de confusión de la SVM lineal con el enfoque de clasificación 5C y utilizando el método de selección de características TFS. 86
4.20. Matriz de confusión de la SVM lineal con el enfoque de clasificación por tipos y utilizando el método de selección de características UFS. 86
4.21. Matriz de confusión de la SVM lineal con el enfoque de clasificación por subtipos y utilizando el método de selección de características UFS. 87

B.1. Notación de las características utlizadas. 106
B.2. Características seleccionadas para el enfoque de clasificación con cinco clases. 107

B.3. Características seleccionadas para el enfoque de clasificación de leucocitosen dos etapas: tipo. 108
B.4. Características selecionadas para el enfoque de clasificación de leucocitos en dos etapas: subtipos. 109

–

Resumen

La sangre está compuesta de plasma y células (eritrocitos, leucocitos y plaquetas). Cada componente en la sangre tiene una función muy importante para el desarrollo de la vida humana. Por una parte, los eritrocitos son células encargadas de realizar el transporte de gases. Por otra parte, las plaquetas tienen un papel importante en la coagulación de la sangre. Además, los leucocitos desempeñan un papel importante en el combate de infecciones u otras enfermedades. El motivo principal de realizar un análisis cuantitativo y cualitativo basado en los componentes celulares de la sangre es que éstas son indicadoras de alteraciones precursoras de enfermedades.

El análisis de frotis de sangre periférica proporciona información acerca del número y forma de las células sanguíneas mediante una exploración visual con la asistencia del microscopio. Por otro lado, al realizar la digitalización de frotis sanguíneo en imágenes digitales, es posible aplicar técnicas de procesamiento digital de imágenes y aprendizaje computacional. Estas disciplinas han mostrado gran capacidad para tratar los problemas de identificación y reconocimiento de células sanguíneas.

El conteo y clasificación de células sanguíneas es una tarea monótona y costosa que consume demasiado tiempo. Aunado a esto, los resultados dependen de la experiencia del técnico laboratorista o químico farmacobiólogo. Por estas razones es importante automatizar el conteo y clasificación de las diferentes células sanguíneas, con la finalidad de reducir el tiempo de estas tareas y así los laboratoristas puedan dedicar más tiempo a los casos que exigen un análisis y una evaluación más cuidadosa.

En este trabajo de tesis se implementa una aplicación para realizar la identificación y conteo automatizado de leucocitos y eritrocitos, además de reconocer 5 tipos de leucocitos en imágenes microscópicas de frotis sanguí-

neos. Para lograr este propósito se utilizan técnicas de procesamiento digital de imágenes y reconocimiento de patrones.

Para alcanzar el objetivo, la metodología propuesta se divide en tres etapas. La primera consiste en identificar y contar automáticamente los leucocitos y eritrocitos en imágenes microscópicas de frotis sanguíneo. Para llevar a cabo estas dos tareas es necesario realizar previamente la identificación de los posibles núcleos de los leucocitos y calcular los centroides, debido a que éstos determinan la posición de los leucocitos. Además, es necesario identificar las células sanguíneas, esto es, generar una buena segmentación de células para evitar regiones fusionadas (células sobrepuestas).

La segunda etapa se enfoca en la generación del conjunto de datos y su procesamiento para entrenar algunos modelos de reconocimiento de leucocitos y seleccionar el que proporcione el mejor rendimiento. Finalmente, en la tercera etapa se desarrolla una aplicación Web que integra las dos etapas previas. En ésta los usuarios pueden cargar una imagen de frotis y la aplicación identifica y cuenta automáticamente los leucocitos y eritrocitos presentes. Además, si el usuario lo desea, la aplicación realiza el reconocimiento de los leucocitos identificados.

Para evaluar el conteo de eritrocitos y regiones candidatas a leucocitos se realizó una comparativa de éstos contra el conteo generado por el especialista. Los experimentos mostraron que se logran recuperar correctamente 209 regiones candidatas a leucocitos de un total de 219 registrados por el especialista. Para el caso de conteo de eritrocitos, los experimentos mostraron que el método de conteo al utilizar restricciones de forma y área, contabilizan 5317 eritrocitos de los 5295 registrados por el especialista.

Con la finalidad de realizar una evaluación cuantitativa de los modelos de clasificación y sus características seleccionadas, se obtuvo que el mejor algoritmo de aprendizaje supervisado para reconocer a los leucocitos es una SVM lineal con una exactitud balanceada del 0.94 utilizando sólo 49 de las 193 características originales. Cabe mencionar que este modelo entrenado es utilizado en la aplicación Web desarrollada.

Capítulo 1

Introducción

La sangre está compuesta de plasma y células (eritrocitos, leucocitos y plaquetas). Cada componente en la sangre tiene una función muy importante para el desarrollo de la vida humana. Por una parte, los eritrocitos son células encargadas de realizar el transporte de gases, es decir, transportar oxígeno desde los pulmones a los tejidos y el dióxido de carbono en sentido inverso. Por otra parte, los leucocitos son las células de la sangre encargadas de reconocer y eliminar cualquier agente extraño del organismo [Moraleda Jiménez, 2017]. Además, las plaquetas tienen un papel importante en la coagulación de la sangre [MedlinePlus, 2018].

El análisis de frotis de sangre periférica proporciona información acerca del número y forma de las células sanguíneas mediante una exploración visual con la asistencia del microscopio [Jaime Pérez and Gómez Almaguer, 2005]. El motivo por el cual el diagnóstico cuantitativo y cualitativo basado en los componentes celulares es de gran importancia es que las células sanguíneas son indicadoras de alteraciones precursoras de enfermedades [Diem et al., 2004].

La biometría hemática es el examen de laboratorio de mayor utilidad y más frecuentemente solicitado debido a que brinda información detallada del estado de salud del paciente de acuerdo con tres tipos de células sanguíneas: eritrocitos, leucocitos y plaquetas [SyM, 2018]. La serie eritrocítica consiste en la cuantificación de índices de eritrocitos primarios y secundarios. Por otro lado, los leucocitos son uno de los componentes principales del sistema inmune, ya que desempeñan un papel importante en el combate de infecciones y otras enfermedades. Para los leucocitos, se realizan tres estudios principales: recuento total, recuento diferencial y recuento diferencial de Schilling

[Jaime Pérez and Gómez Almaguer, 2005].

Los métodos tradicionales de conteo y clasificación de células sanguíneas son tareas monótonas que realiza el técnico laboratorista o químico farmacobiólogo a través de inspección visual. Lo anterior implica algunas desventajas como el consumo de tiempo y que la fiabilidad de los resultados depende de la experiencia del laboratorista al clasificar las células. Por esta razón es importante automatizar el conteo y clasificación de las células sanguíneas con el objetivo de reducir tiempo, costos y estrés a los laboratoristas y así puedan dedicarle más tiempo a otros casos clínicos (como anemias, leucemias, etc.) que requieren un grado de evaluación más cuidadoso.

Gracias a que actualmente se cuenta con la posibilidad de capturar imágenes digitales de frotis sanguíneos, han sido desarrollado múltiples proyectos de investigación y algunos han sido llevados al mercado. Uno de los más destacados es CellaVision® DM9600, un analizador hematológico basado en imágenes digitales, que localiza automáticamente las células en un frotis de sangre periférica, realiza la preclasificación leucocitaria y evalúa morfológicamente los eritrocitos. Este equipo tiene un costo aproximado de $\$31,495.00$ USD.

En el mercado se pueden encontrar equipos que automatizan la identificación y conteo de células, incluso hay algunos que integran la preparación del frotis sanguíneo, realizan la preclasificación de leucocitos y evalúan la morfología de los eritrocitos. Pero desafortunadamente no todas las instituciones o laboratorios pueden tener acceso a equipos tan caros y sofisticados.

Las imágenes digitales de frotis sanguíneo pueden ser utilizadas para tareas de identificación, conteo y reconocimiento de células sanguíneas. Por esta razón múltiples investigadores alrededor del mundo se han enfocado en resolver estas tareas. Por ejemplo, un trabajo utiliza el algoritmo de *watershed* controlado por marcadores para extraer simultáneamente eritrocitos y leucocitos [Miao and Xiao, 2018]. Es decir, en este trabajo únicamente se enfoca en la tarea de identificación.

En otro trabajo [Acharya and Kumar, 2018] proponen segmentar a los leucocitos mediante el algoritmo de agrupamiento *k-medoids* y así determinar el conteo de leucocitos. Posteriormente, mediante un análisis granulométrico separan los eritrocitos de los leucocitos. Los eritrocitos obtenidos son contados utilizando un algoritmo de etiquetado y la transformada circular de

Hough.

Por otro lado, [Zheng et al., 2018] presentan un enfoque de aprendizaje auto-supervisado para mejorar la precisión y adaptabilidad de la segmentación de leucocitos. Éste consta de dos módulos principales: la segmentación inicial no supervisada genera un resultado aproximado de la segmentación y el refinamiento de la segmentación supervisada utiliza estos resultados de la segmentación inicial para entrenar a un clasificador SVM y lograr un resultado de segmentación mejorado.

En [Martínez Castro et al., 2014] desarrollaron un sistema computacional capaz de identificar, clasificar y contar leucocitos mediante un clasificador *k-nn* en combinación con la primera métrica de Minkowski y técnicas de procesamiento digital de imágenes. Por otro lado, en [Jiménez Díaz, 2007] se implementó un clasificador de leucocitos fundamentado en la teoría de redes Bayesianas, para lograr esto, realizaron la extracción de características morfológicas a 5 tipos de leucocitos, mediante técnicas de procesamiento digital de imágenes.

La presente tesis tiene como objetivo implementar una aplicación que integre la identificación y conteo de leucocitos y eritrocitos, así como el reconocimiento de 5 tipos de leucocitos (linfocitos, monocitos, eosinófilos, basófilos y neutrófilos) en imágenes digitales de frotis sanguíneo.

Para alcanzar el objetivo, la metodología propuesta se divide en tres fases: en la primera se identifica y cuenta automáticamente los leucocitos y eritrocitos en imágenes microscópicas de frotis sanguíneo. En la segunda, se construye un reconocedor de patrones de leucocitos. Finalmente en la tercera se desarrolla una aplicación que integre las dos fases previas.

1.1. Planteamiento del Problema

De acuerdo a [López-Santiago, 2016], la biometría hemática o citometría hemática, es el examen de laboratorio de mayor utilidad y más frecuentemente solicitado. La biometría hemática brinda información detallada del estado de salud del paciente de acuerdo con tres tipos de células sanguíneas: eritrocitos, leucocitos y plaquetas [SyM, 2018]. La serie eritrocítica consiste en la cuantificación de índices de eritrocitos primarios y secundarios. En dicho examen las plaquetas también son cuantificadas. Para los leucocitos, se rea-

lizan tres estudios principales: recuento total, recuento diferencial y recuento diferencial de Schiling [Jaime Pérez and Gómez Almaguer, 2005].

En el caso del recuento diferencial de leucocitos se obtiene contando 100 leucocitos mediante microscopio utilizando un frotis sanguíneo teñido con colorante. Este método tradicional para realizar el conteo y clasificación de los leucocitos, es un proceso manual (de inspección visual) tedioso, y el tiempo para realizarlo depende de la experiencia del laboratorista. De modo que seguir realizando el método tradicional implica varias desventajas como el consumo de tiempo del laboratorista para analizar un frotis sanguíneo, además la fiabilidad de los resultados depende de la experiencia del laboratorista para clasificar los leucocitos. Aunado a lo anterior, realizar muchos recuentos diferenciales de leucocitos genera fatiga para los laboratoristas.

Por otro lado, el análisis de las imágenes digitales obtenidas del frotis sanguíneo pueden mejorar el entorno de trabajo y aliviar algunas de las cargas asociadas con la realización de un conteo diferencial de leucocitos. Con imágenes de frotis sanguíneos de buena calidad, las técnicas de procesamiento de imágenes pueden localizar los leucocitos, separar imágenes de esas células y realizar una clasificación preliminar del tipo de célula [Linder and Zahniser, 2012].

En años recientes se han desarrollado diferentes proyectos de investigación cuya finalidad es efectuar la identificación y conteo automático de leucocitos y eritrocitos, así como también el reconocimiento de los leucocitos [Acharya and Kumar, 2018, Putzu and Di Ruberto, 2013, Zheng et al., 2018, Sarrafzadeh et al., 2014].

Una particularidad de los estudios previos es que se han enfrentado por separado los problemas de identificación y conteo de leucocitos y eritrocitos, así como el reconocimiento de leucocitos. Por consecuencia, en este proyecto de tesis se pretende implementar una aplicación que integre la identificación y conteo de leucocitos y eritrocitos, así como el reconocimiento de 5 tipos de leucocitos en imágenes digitales de frotis sanguíneo de sangre periférica. Para lograr esto, la metodología propuesta se divide en tres fases: en la primera se identifica y cuenta automáticamente los leucocitos y eritrocitos en imágenes microscópicas de frotis sanguíneo. En la segunda, se construye un reconocedor de patrones de leucocitos. Finalmente, en la tercera se desarrolla una aplicación que integre las dos fases previas.

1.2. Justificación

Los glóbulos blancos o leucocitos son componentes principales del sistema inmune. Estos desempeñan un papel importante en el combate de infecciones y algunas enfermedades, por lo que se consideran células importantes para la vida humana. Por lo anterior, realizar un análisis de estos es vital para proporcionar información útil a los laboratoristas y así realizar el diagnóstico adecuado de enfermedades.

El conteo y clasificación de células sanguíneas es una tarea monótona y costosa que consume demasiado tiempo, y la fiabilidad de los resultados depende de la experiencia del técnico laboratorista o químico farmacobiólogo. Por estas razones es importante automatizar el conteo y clasificación de las diferentes células sanguíneas, con la finalidad de reducir el tiempo de estas tareas y así los laboratoristas puedan dedicarle más tiempo a los casos que exigen un análisis y una evaluación más cuidadosa.

Para realizar las tareas mencionadas arriba, existen algunos sistemas que realizan el análisis automatizado de la morfología celular. Algunos como CellaVision® DM9600, localiza automáticamente las células en un frotis de sangre periférica, realiza la preclasificación leucocitaria y evalúa morfológicamente los eritrocitos. Otro sistema es Sysmex® DI-60, que integra la preparación de frotis con el análisis digital de la morfología celular en una misma plataforma. Desafortunadamente el costo de los sistemas mencionados es elevado y hace que su acceso sea limitado.

Por otro lado, al realizar la digitalización de frotis de sangre periférica de buena calidad en imágenes digitales, es posible aplicar técnicas de procesamiento digital de imágenes y aprendizaje computacional. Estas disciplinas han mostrado gran capacidad para tratar los problemas de identificación y reconocimiento de células sanguíneas en el campo de la medicina. Entonces, se puede minimizar el esfuerzo de los laboratoristas para acceder a la información visual de interés y ahorrar el tiempo en la lectura y análisis de los frotis.

Aún cuando existen trabajos realizados internacionalmente en los que se abordan las tareas de identificación, conteo y clasificación de células sanguíneas por separado, en México el trabajo de investigación y desarrollo tecnológico tratando esta problemática es mínimo. Por las razones anteriores, en

este proyecto de tesis se plantea implementar una aplicación que integre la identificación y conteo de leucocitos y eritrocitos, así como el reconocimiento de 5 tipos de leucocitos en imágenes digitales de frotis sanguíneo de sangre periférica.

1.3. Hipótesis

Con el uso de técnicas de procesamiento digital de imágenes y reconocimiento de patrones es posible realizar la integración de la identificación y conteo de leucocitos y eritrocitos, así como el reconocimiento de 5 tipos de leucocitos en imágenes de frotis sanguíneo.

1.4. Objetivos

1.4.1. Objetivo general

Implementar una aplicación para realizar la identificación y conteo automatizado de leucocitos y eritrocitos, además de reconocer 5 tipos de leucocitos en imágenes microscópicas de frotis sanguíneos.

1.4.2. Objetivos específicos

- Revisar el estado del arte sobre métodos de procesamiento digital de imágenes para la identificación y conteo de leucocitos en imágenes microscópicas de frotis sanguíneo. Además, revisar métodos sobre reconocimiento de patrones de leucocitos.
- Seleccionar e implementar un método para la identificación y conteo automatizado de leucocitos y eritrocitos.
- Seleccionar un algoritmo de aprendizaje supervisado de la biblioteca scikit-learn del lenguaje de programación python para el reconocimiento de 5 tipos de glóbulos blancos (linfocitos, monocitos, eosinófilos, basófilos y neutrófilos).
- Evaluar resultados de la identificación y conteo de leucocitos y eritrocitos.
- Evaluar resultados del reconocimiento de leucocitos.

1.5. Metas

1. Elaboración de un reporte sobre los métodos de procesamiento digital de imágenes para identificar y contar automáticamente leucocitos y eritrocitos en imágenes microscópicas de frotis sanguíneos.

2. Elaboración de un reporte sobre los métodos sobre reconocimiento de patrones de leucocitos.

3. Implementación de un módulo en python para identificar y contar leucocitos y eritrocitos en imágenes de frotis sanguíneo.

4. Implementación de un módulo en python para reconocer 5 tipos de leucocitos (linfocitos, monocitos, eosinófilos, basófilos y neutrófilos) usando un algoritmo de aprendizaje supervisado.

5. Integración de los módulos desarrollados en los dos puntos anteriores para la creación de una aplicación en python que sea capaz de identificar y contar leucocitos y eritrocitos, y también pueda reconocer a los leucocitos en una imagen de frotis sanguíneo.

6. Elaboración de un reporte comparativo del desempeño de conteo de leucocitos y eritrocitos contra los métodos tradicionales.

7. Elaboración de un reporte comparativo del desempeño de reconocimiento de leucocitos contra los métodos tradicionales.

8. Elaboración del documento de tesis.

1.6. Trabajos relacionados

Dada la problemática de realizar el conteo y clasificación de células sanguíneas de manera eficiente, en la literatura se pueden encontrar distintas investigaciones para reducir el tiempo en estas tareas y realizarlas de manera automática. A continuación, se mencionan algunos trabajos relacionados con este proyecto de tesis.

Para abordar el problema de realizar la identificación y conteo de leucocitos en imágenes digitales, [Putzu and Di Ruberto, 2013] presentan un método completo y automático para la identificación de glóbulos blancos a partir

de imágenes microscópicas. Este método está basado en el umbral del espacio de color cian, magenta, amarillo y negro (CMYK, por sus siglas en inglés).

Por otro lado, en [Rezatofighi and Soltanian-Zadeh, 2011] proponen algoritmos de procesamiento de imágenes para reconocer automáticamente cinco tipos de glóbulos blancos en la sangre periférica. Inicialmente, proponen un método basado en la ortogonalización de Gram-Schmidt junto con un algoritmo de contornos activos (*Snake*) para segmentar el núcleo y el citoplasma. Posteriormente, realizan la extracción de características a los leucocitos, y mediante el algoritmo de selección progresiva secuencial (SFS, por sus siglas en inglés) selecciona las características más importantes y finalmente comparan el rendimiento obtenido con algoritmos de aprendizaje automático.

Debido a que los núcleos de los leucocitos son áreas de alto contraste y son relativamente fáciles de segmentar en la imagen de frotis sanguíneo, algunos de los métodos tradicionales segmentan primero la parte del núcleo de la imagen antes de la parte del citoplasma. Por ejemplo, el trabajo de [Huang and Hung, 2012] se enfoca en segmentar únicamente los núcleos y eliminar manchas en las imágenes de frotis sanguíneo. Posteriormente, extraen características al núcleo, reducen características por análisis de componentes principales (PCA, por sus siglas en inglés) y con un algoritmo genético basado en el método de agrupamiento de *k-means* clasifican los 5 tipos de leucocitos. Por otro lado, [Zamani and Safabakhsh, 2006] propone segmentar al núcleo del leucocito, aplicar flujo de gradiente vectorial (GVF, por sus siglas en inglés) a los leucocitos segmentados, y utilizar la envolvente de los núcleos como contorno inicial para un método de contornos activos.

Los métodos basados en aprendizaje computacional como las máquinas de soporte vectorial (SVM, por sus siglas en inglés), redes neuronales artificiales y los métodos no supervisados son ampliamente utilizados para la segmentación y clasificación de las células sanguíneas. [Zhang et al., 2014] propone un método para la segmentación del núcleo y el citoplasma, asimismo realiza un ajuste en el color previo a la segmentación, así la descomposición del espacio de color y el agrupamiento de *k-means* se combinan para realizar la segmentación. Para ahondar más en el tema, en [Tuan Muda and Abdul Salam, 2013] se presenta un análisis comparativo de varios algoritmos de segmentación.

Por otro lado, realizar el conteo de eritrocitos implica una labor muy

grande, debido a que primero se necesita excluir a los leucocitos presentes en la imagen de frotis sanguíneo. Para esto, en [Acharya and Kumar, 2018] proponen extraer a los leucocitos mediante el algoritmo *k-medoids*, y posteriormente mediante un análisis granulométrico separan los eritrocitos de los leucocitos. Los eritrocitos obtenidos los cuentan utilizando un algoritmo de etiquetado y la transformada circular de Hough. En otro trabajo semejante, [Wei and Cao, 2016] proponen una segmentación automática de eritrocitos superpuestos basados en la predicción de semillas. En esta investigación las regiones de los eritrocitos se extraen de forma rápida y precisa, el algoritmo de *k-means* se aplica en la detección de bordes de eritrocitos superpuestos, y presentan un nuevo método de conteo automático que considera a eritrocitos superpuestos.

Teniendo en cuenta los retos que implican realizar una buena segmentación de leucocitos, otro problema surge al intentar extraer de manera simultánea leucocitos y eritrocitos. Para esto en [Miao and Xiao, 2018] se presenta un algoritmo de Watershed controlado por marcadores para tratar de simplificar estas operaciones y reducir el tiempo de cálculo.

En México, [Ruiz Segura, 2016] describe una metodología general para el proceso de identificación y clasificación de leucocitos basada en algoritmos de visión computacional. De manera similar, [Cuevas et al., 2010] presentan un algoritmo para la segmentación, detección y medición de leucocitos en imágenes de frotis sanguíneo usando Sistemas Inmunes Artificiales.

Por otro lado, [Jiménez Díaz, 2007] implementa un clasificador automático de glóbulos blancos usando las redes Bayesianas. Para esto consideran 5 tipos de leucocitos: neutrófilos, linfocitos, monocitos, eosinófilos y basófilos.

1.7. Metodología

El desarrollo del presente proyecto de tesis consta de diversas etapas, las cuales son mostradas en la Figura 1.1.

La primera etapa consiste en identificar y contar automáticamente los leucocitos y eritrocitos en imágenes microscópicas de frotis sanguíneo. Para lograr esto, primero se realiza la identificación de los posibles núcleos de los leucocitos y se calculan los centroides, debido a que éstos determinan la posición de los leucocitos. Después, se realiza el preprocesamiento a las imá-

genes para reducir el ruido y resaltar las regiones de interés. A las imágenes resultantes en el paso anterior se les aplica un algoritmo de segmentación para eliminar el fondo y solo quedarse con células sanguíneas. Algunas células están sobrepuestas una sobre la otra generando regiones segmentadas muy grandes que agrupan varios elementos, es decir, regiones fusionadas. En consecuencia, se realiza un postprocesamiento para separar las regiones fusionadas y generar una mejor segmentación de células sanguíneas. Una vez segmentadas las células sanguíneas se procede a separar los leucocitos y eritrocitos. Para lograrlo se utilizan los centroides de los núcleos y la imagen segmentada de células. Finalmente, se realiza el conteo de los leucocitos y eritrocitos.

La segunda etapa se enfoca en la generación del conjunto de datos y su procesamiento para entrenar modelos de reconocimiento de leucocitos. Para lograr esto, primero se realiza la extracción de características de color, forma y textura al núcleo, citoplasma y toda la región que abarca toda la célula de los leucocitos identificados. Después del paso de extracción de características, al conjunto de datos resultante se le aplica un preprocesamiento de datos para tener una buena calidad en ellos. Con estos datos se crean modelos para la clasificación de leucocitos. Esto es, primero se seleccionan las características que describen mejor a los leucocitos, a continuación se realizan pruebas con algoritmos de clasificación y se selecciona el que proporcione el mejor rendimiento.

En la tercera etapa se desarrolla una aplicación usando el lenguaje de programación python para integrar y mostrar el uso práctico de esta metodología. Esta aplicación recibirá una imagen de frotis sanguíneo, identifica y cuenta automáticamente los leucocitos y eritrocitos. A continuación si el usuario lo desea, la aplicación realiza la extracción de las características a cada uno de los leucocitos identificados y el reconocimiento para determinar el tipo de leucocito de cada una de las células identificadas.

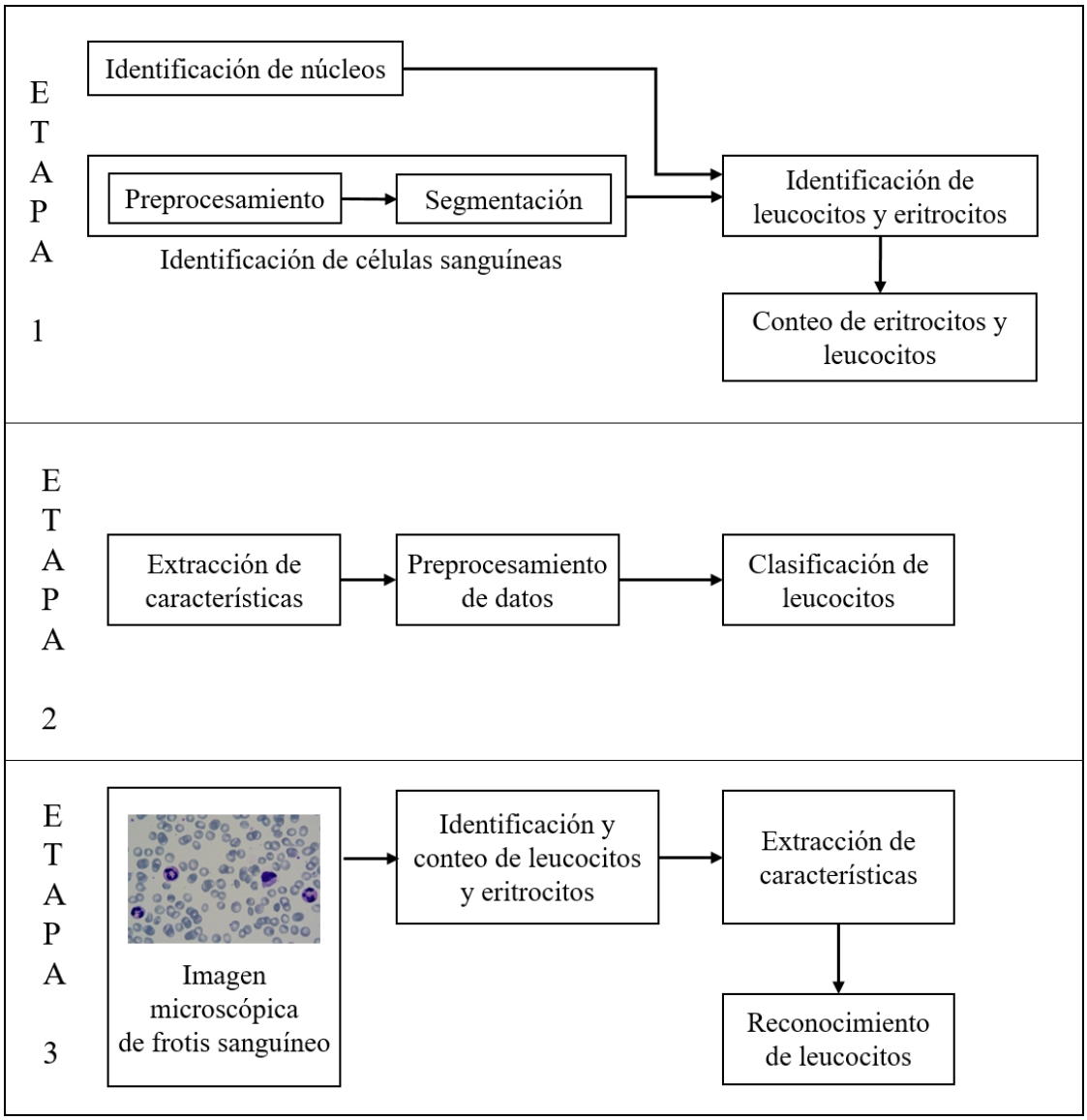


Figura 1.1: Metodología para la identificación, conteo de leucocitos y eritrocitos, asimismo para el reconocimiento de 5 tipos de leucocitos.

Capítulo 2

Marco Teórico

En este capítulo se describe en la primera sección el frotis sanguíneo y las células presentes en la sangre. En la segunda sección lo que es el Procesamiento Digital de Imágenes, sus etapas y se presentan los métodos para realizar el preprocesamiento de imágenes de frotis sanguíneo. Asimismo, se muestran las técnicas utilizadas para identificar y contar los leucocitos y eritrocitos en imágenes de frotis. Por otro lado, en la tercera sección, se explican los componentes de un Sistema de Reconocimiento de Patrones. Además, se presentan las características de forma, textura y color que describen a los leucocitos. También, se describe cada uno de los algoritmos de aprendizaje supervisado utilizados para poder realizar el reconocimiento de los leucocitos. Finalmente, en la cuarta sección se describen algunos enfoques de clasificación de leucocitos.

2.1. Frotis y células sanguíneas

Para una mejor comprensión sobre el frotis sanguíneo, es necesario primero conocer a las células sanguíneas.

La sangre está constituida por un líquido denominado plasma y tres clases de células, cada una de las cuales desempeña una función específica. La célula denominada glóbulo blanco o leucocito es la principal defensa del cuerpo contra las infecciones y las sustancias extrañas que pudieran entrar en él. Al igual que todas las células sanguíneas, los leucocitos son: producidos en la médula ósea. Los cinco tipos principales de leucocitos son neutrófilos, linfocitos, monocitos, eosinófilos y basófilos. Otra célula importante es el glóbulo

rojo, también llamado eritrocito que se ocupa de transportar el oxígeno desde los pulmones a los tejidos, y de llevar de vuelta el dióxido de carbono de los tejidos hacia los pulmones para su expulsión. Por otra parte, la célula denominada plaqueta o trombocito colabora en la coagulación de la sangre cuando se produce la rotura de un vaso sanguíneo.

Una vez conocidas las células sanguíneas, se define el frotis como una gota de sangre extendida sobre un portaobjetos y teñida con un colorante. Éste permite el estudio de la morfología de los eritrocitos y las alteraciones de su color y tamaño. En el frotis también se observa la morfología de los leucocitos y las plaquetas. Para mayor información sobre las células sanguíneas y el frotis ver Anexo A.

2.2. Procesamiento Digital de Imágenes

El procesamiento digital de imágenes puede definirse como un conjunto de procedimientos que se realizan sobre una imagen digital y sus resultados son imágenes y además, incluye procesos que extraen atributos de imagen y también realizan el reconocimiento de objetos [Gonzalez and Woods, 2008]. Debido a la falta de un acuerdo general sobre dónde termina el procesamiento de imágenes y dónde comienzan otras áreas relacionadas, a continuación se describe una categorización de procesos que gozan de consenso en la comunidad. El primer proceso es denominado de bajo nivel, debido a que se realiza el preprocesamiento de imágenes para reducir el ruido, mejorar el contraste y el enfoque en la imagen. Este proceso se caracteriza porque sus entradas son imágenes y las salidas también. El segundo proceso es denominado de nivel medio porque se realizan tareas de segmentación de objetos, descripción de esos objetos para el procesamiento por la computadora y así clasificarlos. El nivel medio se caracteriza porque sus entradas son imágenes pero sus salidas son atributos extraídos de la imagen. El proceso de alto nivel implica realizar una interpretación del conjunto de objetos reconocidos, esto se entiende como una etapa superior como el análisis de imágenes o ejecutar funciones cognitivas asociadas con la visión. Este proyecto de tesis se relaciona con los dos primeros niveles.

2.2.1. Etapas del Procesamiento Digital de Imágenes

[Gonzalez and Woods, 1996] transmiten una idea de las diferentes técnicas que se pueden aplicar a las imágenes para diferentes propósitos. De acuerdo a la Figura 2.1, el procesamiento digital de imágenes se compone de las siguientes etapas, donde un sistema en particular puede o no contener todas ellas.

- Preprocesamiento: Mejorar la imagen de forma que se incrementen las posibilidades de éxito en los procesos posteriores.
- Segmentación: Consiste en dividir una imagen de entrada en regiones de interés.
- Representación y descripción: Consiste en extraer características con alguna información cuantitativa de interés o que sean cruciales para reconocer un objeto de otro.
- Reconocimiento: Proceso que asigna una etiqueta a un objeto basándose en la información proporcionada por sus descriptores.
- Interpretación: Implica asignar un significado a un conjunto de entidades etiquetadas.

Con respecto a las etapas posteriores a segmentación, estas se abordarán más detalladamente en la Seccion 2.3

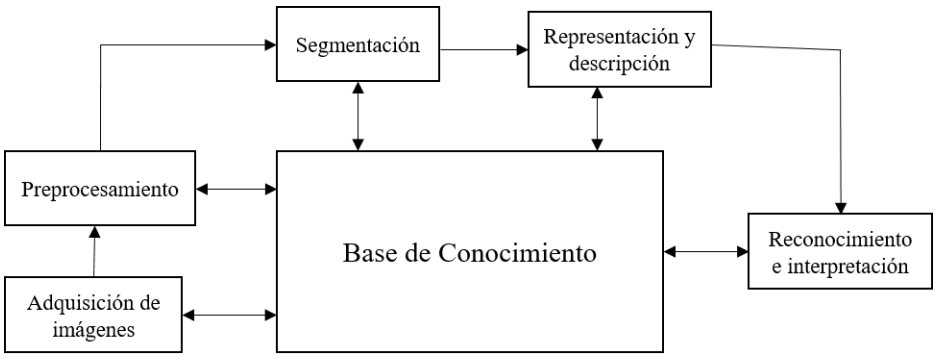


Figura 2.1: Etapas del Procesamiento Digital de Imágenes.

2.2.2. Preprocesamiento de imágenes digitales de frotis sanguíneo

En este proyecto de tesis los métodos de preprocesamiento de imágenes se utilizan para reducir el ruido y resaltar las regiones de interés en las imágenes de frotis sanguíneo.

Conversión de color a escala de grises

Las imágenes de frotis sanguíneo se convierten a escala de grises para eliminar el matiz (*hue*) y saturación en las imágenes, y conservar luminancia [Acharya and Kumar, 2018]. La escala de grises es la representación de una imagen en tonos de grises, es decir un arreglo matricial de dos dimensiones que aporta información sobre la intensidad de la luz presente para cada punto de la imagen [Burger and Burge, 2009]. Entonces, para calcular la luminancia y así realizar la conversión a escala de grises, ésta se define como la suma ponderada de los canales rojo, verde y azul de una imagen:

$$Y = 0.2125 * R + 0.7154 * G + 0.0721 * B \tag{2.1}$$

donde Y es la luminancia de una imagen en el espacio de color RGB. Si hay un canal alfa presente en la imagen, se ignora [Poynton, 1997].

Transformación gamma

La corrección gamma se basa en la función de potencia:

$$s = cr^{\gamma} \tag{2.2}$$

donde c y γ son constantes positivas, y el valor de los píxeles, antes y después de la transformación, son denotados como r y s, respectivamente. Además, en la práctica el valor de c es fijado a 1 generalmente. Al realizar el ajuste de intensidades se busca mapear los valores de intensidad de la imagen a otros nuevos. Por ejemplo, para valores de $\gamma < 1$ la imagen se aclara, es decir, la mayor parte del rango de valores de entrada se mapea a valores altos del rango de valores de salida, con $\gamma = 1$ no hay cambios en los valores de las intensidades. Para $\gamma > 1$ la imagen se oscurece, es decir, la mayor parte del rango de valores de entrada se mapea a valores bajos del rango de valores de salida [Gonzalez and Woods, 2008].

Filtro mediana

El filtro mediana es uno de los filtros espaciales no lineales debido a que se basa en ordenar los píxeles contenidos en la vecindad, y luego reemplaza el valor del píxel central por el valor de la mediana. De este modo, el filtro mediana remplaza el valor de cada píxel en la imagen por la mediana de los valores de intensidad en su vecindad. Cabe mencionar que el valor original del píxel es incluido en el cálculo de la mediana [Gonzalez and Woods, 2008].

Ecualización adaptativa del histograma limitada en contraste

Es un algoritmo para la mejora de contraste local, que utiliza histogramas calculados en diferentes regiones denominadas azulejos (tildes) de la imagen. De este modo, los detalles locales se pueden mejorar incluso en regiones que son más oscuras o más claras que la mayoría de la imagen. Además el contraste, especialmente en áreas homogéneas, puede limitarse para evitar cualquier amplificación de ruido que pueda presentarse en la imagen [Zuiderveld, 1994].

2.2.3. Técnicas utilizadas para identificar leucocitos y eritrocitos

En esta subsección se presentan algunas técnicas de Procesamiento Digital de Imágenes que son utilizadas para realizar la identificación de los leucocitos y eritrocitos, asimismo se hace mención de los métodos utilizados para la identificación de los núcleos de los leucocitos.

Umbralización

La umbralización o binarización consiste en convertir una imagen en escala de grises a una imagen de solo dos intensidades. Una forma de lograr esto es calcular un umbral, T, para separar las regiones u objetos de interés y el fondo de la imagen [Ávarez et al., 2006]. Así la imagen segmentada $g(x, y)$ está dada por:

$$g(x,y) = \begin{cases} 1, & \text{Si } f(x,y) > T \\ 0, & \text{Si } f(x,y) \leq T \end{cases} \tag{2.3}$$

Cuando el valor de T es una constante que se aplica a la imagen completa, se habla de una umbralización global. Cuando el valor de T en cualquier punto

(x, y) en la imagen depende de los atributos de una vecindad, se habla de una umbralización variable [Gonzalez and Woods, 2008].

Umbralización global de Otsu

El método de Otsu elige un umbral óptimo maximizando la varianza entre dos clases [Otsu, 1979], es decir, un umbral que dé la mejor separabilidad entre clases en términos de sus valores de intensidad. Entonces el valor del umbral óptimo k^* que maximiza es:

$$\sigma_B^2(k^*) = \max_{0 \leq k \leq L-1} \sigma_B^2(k) \tag{2.4}$$

donde L denota niveles de intensidad en una imagen, $\sigma_B^2(k)$ mide la varianza intra clases y está es definida como:

$$\sigma_B^2(k) = \frac{[m_G P_1(k) - m(k)]^2}{P_1(k)\,[1 - P_1(k)]} \tag{2.5}$$

donde $P_1(k)$ es la probabilidad de ser asignado a la primera clase C_1 (de las dos divididas por el umbral), m_G es la media global y $m(k)$ es la media acumulativa.

Umbralización de Yen

En [Yen et al., 1995] proponen un nuevo criterio para la umbralización multinivel. Este criterio considera dos factores, el primer factor es la discrepancia entre la umbralización y la imagen original, el segundo factor es el número de bits requeridos para representar la imagen umbralizada. Para definir esa discrepancia, primero hacen el análisis del criterio de máxima entropía (MEC, por sus siglas en inglés) y después proponen el criterio de máxima correlación para la umbralización en dos niveles. En este proyecto de tesis se utiliza principalmente el primer factor.

Sea $f(x, y)$ una imagen de $N \times N$ píxeles con m niveles de intensidades en gris. $G_m = \{0, 1, ...(m-1)\}$ denota el conjunto de niveles de intensidad y f_i las frecuencias de nivel observadas en la imagen. La probabilidad del nivel de intensidad i en la imagen se calcula:

$$p_i = \frac{f_i}{N \times N},\ i \in G_m \tag{2.6}$$

De este modo, una distribución $\{p_i \mid i \in G_m\}$ puede ser obtenida. Así para un nivel de intensidad s se obtienen dos distribuciones

$$A \equiv \left\{ \frac{p_0}{P(s)}, \frac{p_1}{P(s)} ... \frac{p_{s-1}}{P(s)} \right\} \tag{2.7}$$

$$B \equiv \left\{ \frac{p_s}{1-P(s)}, \frac{p_{s+1}}{1-P(s)}, ..., \frac{p_{m-1}}{1-P(s)} \right\} \tag{2.8}$$

donde $P(s) = \sum_{i=0}^{s-1} p_i$ es la probabilidad acumulada. Entonces, se puede definir la correlación total de las distribuciones A y B como:

$$\begin{aligned} TC(s) &= C_A(s) + C_B(s) \\ &= -\ln \sum_{i=0}^{s-1} \left(\frac{p_i}{P(s)} \right)^2 - \ln \sum_{i=s}^{m-1} \left(\frac{p_i}{1-P(s)} \right)^2 \end{aligned} \tag{2.9}$$

Al utilizar la fórmula 2.9, para obtener la máxima correlación entre los objetos y el fondo de la imagen, se determina un umbral s^* tal que:

$$TC(s^*) = \max_{s \in G_m} TC(s) \tag{2.10}$$

Transformación de Watershed

La transformación Watershed o también conocida como cuencas hidrográficas se basa en visualizar una imagen en 3 dimensiones: 2 coordenadas espaciales y el nivel de intensidad. En cuanto a la interpretación topográfica, ver Figura 2.2 , se consideran 3 tipos de puntos:

1. Puntos que pertenecen a un mínimo local.

2. Los puntos en los que al colocar agua caen en un mínimo local. Al cumplir dicha condición son llamados cuencas de captación o watershed.

3. Los puntos en los que el agua cae con igual probabilidad en más de un mínimo local, son llamados líneas de watershed.

La idea general de Watershed es suponer que se han perforado todos los mínimos locales y toda la superficie comienza a inundarse desde abajo,

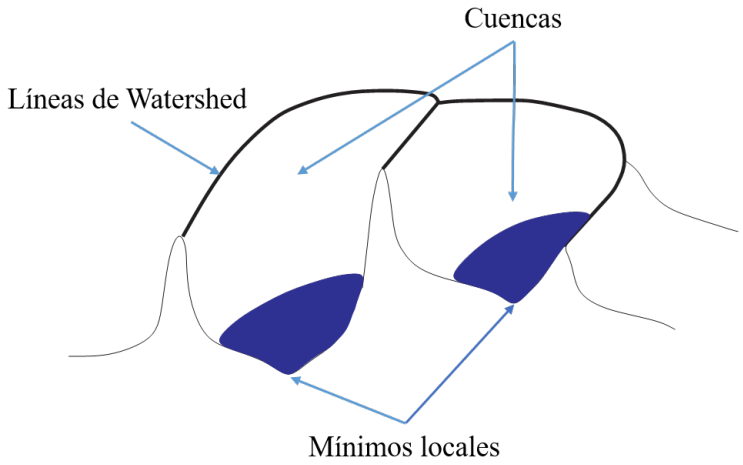


Figura 2.2: Minimos Locales, Cuencas y Lineas de Watershed.

permitiendo que el agua fluya a través de los agujeros de manera uniforme. Cuando el agua fluye por distintas cuencas de captación que están a punto de fusionarse se construye una presa para evitarlo. La inundación continúa hasta que solo las cimas de las presas son visibles. Los límites de las presas corresponden a las líneas de Watershed. Por lo que esas líneas son los límites entre las regiones [Gonzalez and Woods, 2008].

Erosión y dilatación

La erosión y dilatación son operaciones morfológicas, las cuales están basadas en términos de las operaciones básicas de conjuntos. Las operaciones son realizadas con A una imagen y B un elemento estructurante. La erosión tiende a hacer las regiones más pequeñas removiendo pixeles, mientras que la dilatación tiende a verse visualmente como una expansión de las regiones existentes en la imagen.

La erosión de A por B es denotado $A \ominus B$ y se define como:

$$A \ominus B \{z \mid (B)_z \subseteq A\} \tag{2.11}$$

donde $(B)_z$ es la traslación de B por el punto $z = (z_1, z_2)$, es decir $(B)_z$ es el conjunto de puntos en B, cuyas coordenadas (x, y) han sido remplazados por $(x + z_1, x + z_2)$.

La dilatación de A por B es denotado $A \oplus B$ y se define como:

$$A \oplus B = \left\{ z \mid (\hat{B})_z \cap A \neq \emptyset \right\} \tag{2.12}$$

donde $(\hat{B})_z$ representa el conjunto de puntos en B, cuyas coordenadas (x, y) ha sido remplazados por $(-x, -y)$ y después se realiza la traslación por el punto $z = (z_1, z_2)$.

Cierre y apertura

Las operaciones morfológicas de erosión y dilatación se combinan dando como resultado dos operaciones morfológicas. La primera es la operación de apertura, la cual suaviza el contorno de una región, rompe franjas estrechas y elimina protuberancias. La segunda es la operación de cierre que generalmente fusiona las roturas, elimina pequeños orificios y rellena huecos en el contorno.

La apertura $A \circ B$ se define como la erosión de A y B, seguido de una dilatación del resultado por B, de la siguiente manera:

$$A \circ B = (A \ominus B) \oplus B \tag{2.13}$$

El cierre $A \bullet B$ es la dilatación de A y B, seguido de una erosión del resultado por B y se define de la siguiente manera:

$$A \bullet B = (A \oplus B) \ominus B \tag{2.14}$$

Rellenado de agujeros

En [Gonzalez and Woods, 2008], presentan un algoritmo basado en la dilatación, intersección y el complemento para rellenar agujeros en la imagen, el cual está definido de la siguiente manera:

$$X_k = (X_{k-1} \oplus B) \cap A^c \quad k = 1, 2, 3... \tag{2.15}$$

donde A es una imagen con regiones que contienen agujeros, A^c es el complemento de la misma y B es un elemento estructurante simétrico. Al inicio X_0 es una matriz del mismo tamaño que A inicializada con 0s, excepto en las ubicaciones que corresponden a un punto conocido como agujero, los cuales son fijados a 1. El algoritmo termina cuando $X_k = X_{k-1}$. De este modo X_k contiene todos los agujeros rellenados y la unión de X_k y A contiene todas las regiones sin agujeros.

2.2.4. Técnicas utilizadas para contar leucocitos y eritrocitos

Para realizar el conteo de leucocitos es necesaria la extracción de componentes conectados. Entonces para los eritrocitos se suele combinar la extracción de componentes conectados con el criterio para identificar eritrocitos normales.

Extracción de componentes conectados

Sea A una imagen binaria con un conjunto de uno o más componentes conectados. Además, X_0, una matriz del mismo tamaño que A, cuyos elementos son inicializados con cero, excepto en los puntos que se conoce forman parte de una componente, a los cuales se les asigna 1. El objetivo es comenzar con X_0 y encontrar todos los componentes conectados. Este proceso iterativo está definido de la siguiente manera:

$$X_k = (X_{k-1} \oplus B) \cap A \quad k = 1, 2, 3... \tag{2.16}$$

donde B es un elemento estructurante, de este modo el proceso termina cuando $X_k = X_{k-1}$, con X_k almacenando la información de todos los componentes conectados de la imagen de entrada.

Factor de forma

En [Acharya and Kumar, 2018] proporcionan un criterio para identificar eritrocitos normales, el criterio se llama factor de forma y se define como:

$$FactordeForma = \frac{4\pi \text{Área}}{\text{Perímetro}^2} \tag{2.17}$$

Si el factor de forma se encuentra en el rango de $(0.5, 1)$ entonces se considera como una célula normal, de lo contrario se define como anormal.

2.3. Reconocimiento de Patrones

Debido a que la definición de Reconocimiento de Patrones ha tenido múltiples versiones, en este trabajo adoptaremos la definición mostrada en [Bishop, 2006]. Aquí, el campo de reconocimiento de patrones se relaciona

con el descubrimiento automático de regularidades en los datos mediante el uso de algoritmos computacionales y con el uso de estas regularidades para tomar acciones tales como clasificar los datos en diferentes categorías.

En esta sección se describen los componentes de un sistema de reconocimiento de patrones. Después, se explican cada una de las características de forma, textura y color utilizados para describir una región. Posteriormente, se explica por qué se realiza normalización al conjunto de datos obtenidos en la etapa de extracción de características y se describen los métodos de selección de características. Finalmente, se hace mención de los algoritmos de clasificación utilizados para realizar el reconocimiento de leucocitos.

2.3.1. Componentes de un Sistema de Reconocimiento de Patrones

Un Sistema de Reconocimiento de Patrones consta de diversos componentes, los cuales son mostrados en la Figura 2.3 [Duda et al., 2000] y descritos a continuación.

- Sensado: La entrada de un sistema de reconocimiento de patrones, frecuentemente, es un transductor. Este sensor convierte entradas físicas en datos con señales digitales.

- Segmentación y agrupación: Este componente aísla los objetos sensados del fondo o de otros objetos. Esta tarea es dependiente del dominio y del problema.

- Extracción de características: El objetivo es caracterizar un objeto a ser reconocido a través de medidas, las cuales son muy similares para objetos en la misma categoría y muy diferentes para objetos de diferentes categorías.

- Clasificación: El objetivo de este componente es utilizar un vector de variables proporcionado por el extractor de características para asignar el objeto a una categoría.

- Post-procesamiento: Este componente se usa para realizar una acción recomendada en base a la salida del clasificador. Estas acciones deben considerar minimizar el costo total esperado (el riesgo). El post-

procesador debe ser capaz de tomar información del contexto del problema y mejorar el rendimiento del sistema.

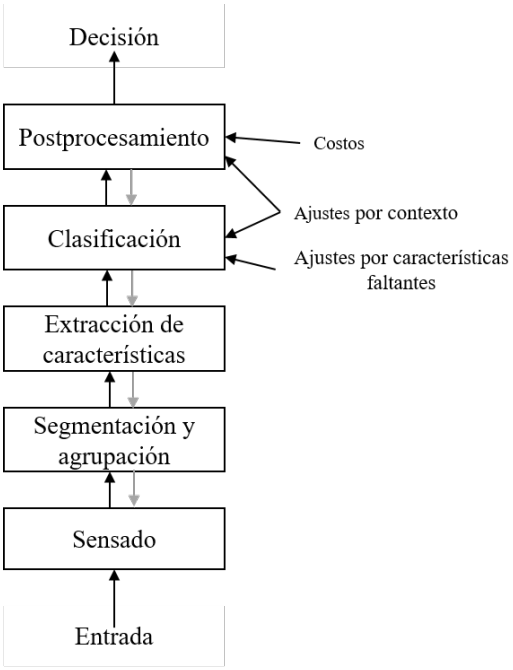


Figura 2.3: Componentes de un Sistema de Reconocimiento de Patrones.

2.3.2. Extracción de características

La representación de una región se puede realizar de dos formas, la primera en términos de sus puntos interiores (los píxeles que comprenden la región) y la segunda en términos de sus puntos exteriores (pixeles del contorno) [Gonzalez and Woods, 2008].

Descriptores de región básicos

Una región puede ser descrita considerando medidas escalares basadas en sus propiedades geométricas como se definen en [Nixon and Aguado, 2012]. Los descriptores utilizados en esta tesis son los siguientes.

Área: Es el número de píxeles contenidos dentro de lo límites de la región

S y se define como:

$$A(S) = \sum_x \sum_y I(x, y)\Delta A \tag{2.18}$$

donde ΔA es el área de un píxel. Entonces si $\Delta A = 1$, el área se mide en píxeles.

Perímetro: El número de píxeles que se encuentran en el contorno de la región y se define como:

$$P(S) = \sum_i \sqrt{(x_i - x_{i-1})^2 + (y_i - y_{i-1})^2} \tag{2.19}$$

donde x_i y y_i representan las coordenadas del i-ésimo píxel que forma parte de la curva (contorno). Tomar en cuenta que la aproximación discreta en la Equación 2.19 produce pequeños errores en el perímetro calculado.

Compacidad: Dado el área y el perímetro de la región, la compacidad se define como:

$$C(S) = \frac{4\pi A(s)}{P^2(s)} \tag{2.20}$$

Dispersión: Es la razón de la cuerda de mayor longitud de la región entre el área de la misma [Chen et al., 1995]. De este modo, la dispersión describe la densidad de la región.

$$I(S) = \frac{\pi \,\text{máx}((x_i - \bar{x})^2 + (y_i - \bar{y})^2)}{A(S))} \tag{2.21}$$

donde $(\bar{x}, \bar{y})$ representan las coordenadas del centro de masa de la región. Una medida alternativa de la dispersión puede expresarse también como la relación entre el radio máximo y el mínimo, es decir, una forma alternativa de la irregularidad.

$$IR(S) = \frac{\text{máx}(\sqrt{(x_i - \bar{x})^2 + (y_i - \bar{y})^2})}{\text{mín}(\sqrt{(x_i - \bar{x})^2 + (y_i - \bar{y})^2})} \tag{2.22}$$

Esta medida define la relación entre el radio del círculo máximo que rodea la región y el círculo máximo que puede contener la región.

Momentos invariantes de Hu

Los momentos de una imagen describen la distribución de los píxeles en una región. El momento de orden $(p+q)$ para una imagen digital $f(x, y)$ de tamaño $M * N$ se define de la siguiente manera en [Gonzalez and Woods, 2008]:

$$m_{pq} = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} x^p y^q f(x, y) \tag{2.23}$$

donde $p = 0, 1, 2, ...$ y $q = 0, 1, 2, ...$ son enteros. El momento central de orden $(p + q)$ se define como:

$$\mu_{pq} = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} (x - \bar{x})^p (y - \bar{y})^q f(x, y) \tag{2.24}$$

para $p = 0, 1, 2, ...$ y $q = 0, 1, 2, ...$, donde

$$\bar{x} = \frac{m_{10}}{m_{00}} \quad \text{y} \quad \bar{y} = \frac{m_{01}}{m_{00}} \tag{2.25}$$

Los momentos centrales normalizados, denotado η_{pq}, se definen como:

$$\eta_{pq} = \frac{\mu_{pq}}{\mu_{00}^{\gamma}} \tag{2.26}$$

donde

$$\gamma = \frac{q + p}{2} + 1 \tag{2.27}$$

para $p + q = 2, 3, ...$

Los siete momentos invariantes de Hu se pueden obtener usando únicamente los momentos centrales normalizados de orden 2 y 3, los cuales se definen en [Hu, 1962] y los cuatro primeros son:

$$\phi_1 = \eta_{20} + \eta_{02} \tag{2.28}$$

$$\phi_2 = (\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2 \tag{2.29}$$

$$\phi_3 = (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2 \tag{2.30}$$

$$\phi_4 = (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2 \tag{2.31}$$

Momentos de Zernike

Los momentos de Zernike tienen muchas propiedades, como invarianza a rotación, robustez al ruido, eficiencia de expresión, rapidez y representación multinivel para describir formas de patrones [Kim and Kim, 2000].

Los Momentos de Zernike, definidos en [Nixon and Aguado, 2012] como Z_{pq}, se expresan en términos de polinomios por la ecuación:

$$Z_{pq} = \frac{p+1}{\pi} \int_0^{2\pi} \int_0^{\infty} V_{pq}(r,\theta)^* f(r,\theta) dr\, d\theta \tag{2.32}$$

donde p es una magnitud radial, q es la dirección radial y $V_{pq}(r,\theta)^*$ denota el complejo conjugado del conjunto de polinomios de Zernike definido por

$$V_{pq}(r,\theta) = R_{pq}(r) e^{jq\theta} \text{ donde } p-q \text{ es par y } 0 \leq q \leq \mid p \mid \tag{2.33}$$

R_{pq} es un polinomio de valor real dado por:

$$R_{pq} = \sum_{m=0}^{\frac{p-|q|}{2}} (-1)^m \frac{(p-m)!}{m! \left(\frac{p-|q|}{2} - m\right)! \left(\frac{p-|q|}{2} - m)\right)!} r^{p-2m} \tag{2.34}$$

El orden del polinomio se denota por p y la repetición por q.

Descriptores de textura de Haralick

En [Haralick et al., 1973] proponen un conjunto de medidas estadísticas basados en la textura como un patrón de niveles de grises. Esta técnica ha sido ampliamente utilizada en aplicaciones de análisis de imágenes, en el área de biomedicina [Zayed and Elnemr, 2015]. Estas medidas de textura son calculadas en base a la matriz de co-ocurrencia de niveles de grises (GLCM, por sus siglas en inglés). La GLCM es denotada como $P(i,j;d,a)$ y representa la probabilidad de ocurrencia de un par de niveles de intensidad (i,j) separados por una distancia d y un ángulo a. GLCM se forma de distancias de pixeles unitarias y ángulos de $0°, 45°, 90°$ y $135°$. Esta matriz tiene dimensión cuadrada N_g, donde N_g es el número de niveles de gris de la imagen.

Los descriptores de textura que se extraen a partir de la matriz GLCM se describen en [Haralick et al., 1973].

Descriptores de color

Los descriptores de color fueron considerados en términos del trabajo de [Isaza et al., 2018], el cual consiste en extraer los componentes de los espacios de color RGB, YCbCr y HSV. En este trabajo se agrega también el espacio de color L*a*b* a la imagen y así calcular la media μ y varianza σ^2 de cada componente en cada espacio de color como descriptores.

2.3.3. Reconocimiento de leucocitos

Para el reconocimiento de leucocitos, primero es necesario preprocesar los datos de las variables calculadas. Después, seleccionar las más relevantes y que aportan mayor información para discriminar finalmente las clases de leucocitos consideradas.

Normalización de datos

La normalización de los datos es útil para evitar que las características con valores grandes tengan mayor influencia en la función de costo para el diseño del clasificador. La normalización restringe los valores de todas las características dentro de rangos predeterminados [Theodoridis and Koutroumbas, 2008]. Sean min_X y max_X el valor mínimo y máximo respectivos y calculados utilizando los valores de una característica específica, X. La normalización se define como:

$$\hat{x_i} = scale * x_i + min - min_X * scale, \quad scale = \frac{max - min}{max_X - min_X} \tag{2.35}$$

donde $\hat{x_i}$ es el valor normalizado, $min = 0$ y $max = 1$.

Selección de características

Los objetivos de la selección de variables son mejorar el rendimiento de predicción, hacer que los procesos de cálculo sean más rápidos y proporcionar una mejor comprensión del proceso que genera los datos.

A continuación, se describen algunos métodos para la selección de variables [Guyon and Elisseeff, 2003].

***Ranking* de Variables:** Los algoritmos de selección de variables regularmente incluyen el *ranking* de variables debido a que estos son simples y

escalables. Para el *ranking* de variables se considera un conjunto de m ejemplos $\{x_k, y_k\}\,(k = 1, ...m)$ que tienen n variables de entrada $x_{k,i}\,(i = 1, ...n)$ y una variable de salida y_k. El *ranking* de variables hace uso de una función de puntuación $S(i)$ calculada a partir de los valores $x_{k,i}$ y y_k, $k = 1, ...m$. Se supone que un puntaje alto es indicativo de una variable valiosa y se ordenan en orden decreciente de $S(i)$.

El *ranking* de variables es considerado un método de filtro y además su complejidad computacional es menor comparando con otros métodos.

Método empaquetador(Wrappers): La metodología *wrapper* ofrece una manera simple y poderosa de abordar el problema de la selección de un subconjunto de variables, independientemente de la máquina de aprendizaje elegida ya que ésta se considera una caja negra perfecta. Esta metodología consiste en utilizar el rendimiento de predicción de una máquina de aprendizaje dada para evaluar la utilidad de subconjuntos de variables. Para utilizar esta metodología se debe definir lo siguiente:

1. Cómo buscar el espacio de todos los posibles subconjuntos de variables

2. Cómo evaluar el rendimiento de predicción de una máquina de aprendizaje para guiar la búsqueda y detenerla

3. Qué predictor usar.

Se puede realizar una búsqueda exhaustiva, si el número de variables no es demasiado grande. Pero, este problema es NP-duro y la búsqueda se puede volver rápidamente intratable computacionalmente. Las evaluaciones de rendimiento usualmente se realizan usando un conjunto de validación o mediante validación cruzada. Los predictores populares incluyen árboles de decisión, clasificadores Bayesianos, predictores lineales de mínimos cuadrados y máquinas de soporte vectorial.

En este proyecto, se utilizaron varios métodos de selección de características, incluyendo la eliminación de características recursivas (RFE, por sus siglas en inglés), selección de características basada en un umbral (TFS, por sus siglas en inglés) y la selección de características univariadas (UFS, por sus siglas en inglés).

Selección de características univariadas: toma las mejores características basándose en una prueba estadística univariada. Hay muchas opciones

diferentes para la selección univariante, para este proyecto de investigación elegimos como función de puntuación la información mutua (MI, por sus siglas en inglés). Ésta mide la dependencia entre las variables y la etiqueta [Wei and Stocker, 2016] para seleccionar las k mejores características.

Eliminación recursiva de características: La eliminación de características recursivas es una instancia de eliminación de características hacia atrás. En tal caso, el método produce un ranking de subgrupos de características, a diferencia de un ranking de características. Se propuso inicialmente para permitir que las máquinas de soporte vectorial realizaran la selección de características entrenando de forma iterativa un modelo, clasificando las características, y luego eliminando las características de clasificación más bajas [Guyon et al., 2002]. Este método se ha aplicado de manera similar a los Bosques Aleatorios [Darst et al., 2018] y se ha comprobado que es beneficioso en presencia de características correlacionadas [Gregorutti et al., 2017].

Selección de características basada en un umbral: La importancia de las características seleccionadas se calcula usando el algoritmo de árboles extremadamente aleatorios (*Extremely Randomized Trees*). Entonces, las características se seleccionan con una importancia mayor o igual al valor de umbral t. De este modo, aquellas características que sean inferiores al valor umbral se eliminan debido a que su importancia es menor.

Clasificación

Usando un enfoque de Aprendizaje Computacional se han desarrollado diversos modelos para la resolución de problemas de clasificación, algunos de ellos se describen a continuación.

Máquinas de Soporte Vectorial: Las máquinas de soporte vectorial (SVM, por sus siglas en inglés) se caracterizan por mapear los puntos de entrada a un espacio de características de alta dimensión, mediante el uso de funciones como los kernels, de manera que se encuentre un hiperplano que separe las clases [Vapnik, 1998]. Existen 4 kernels básicos en la literatura, estos son: kernel lineal, polinomial, función de base radial (RBF, por sus siglas en inglés) y sigmoide. El kernel RBF mapea ejemplos no lineales en un espacio de alta dimensión, por lo que puede manejar el caso cuando las etiquetas de clase y los atributos no tienen una relación lineal [Burges, 1998, Vapnik, 1998].

Bayes Ingenuo: es un clasificador probabilístico basado en aplicar la regla de Bayes pero asumiendo que sus características son independientes estadísticamente [Dougherty, 2013]. Este clasificador asigna la clase más probable (probabilidad más alta) a un ejemplo (patrón desconocido).

***k-nn*:** encuentra un grupo de k ejemplos en el conjunto de entrenamiento que están más cerca del ejemplo de prueba y le asigna la clase que predomina en ese vecindario. Es decir, para clasificar un patrón desconocido se calcula alguna métrica de distancia de ese mismo a todos los elementos del conjunto de entrenamiento, se identifica sus k-vecinos más cercanos y las clases de sus vecinos se usan para determinar la clase del patrón desconocido [Wu et al., 2008].

***Árbol de decisión*:** es un clasificador simple con estructura jerárquica en forma de árbol, que realiza una clasificación supervisada utilizando la estrategia "divide y vencerás" [Dougherty, 2013]. Un árbol de decisión ejecuta una secuencia de características de prueba, iniciando con el nodo raíz y el resultado se obtiene cuando se alcanza un nodo hoja.

Los árboles de decisión son generalmente procesos recursivos. En cada paso, se da un conjunto de datos y se selecciona una división, luego esta división se utiliza para dividir el conjunto de datos en subconjuntos, y cada subconjunto para el siguiente. La clave de este algoritmo es como seleccionar estas divisiones [Zhou, 2012]. A través de los años han surgido diferentes propuestas para seleccionar estas divisiones tales como el algoritmo ID3 [Quinlan, 1986] el cual emplea la entropía como criterio de ganancia para seleccionar esas divisiones. CART [Breiman et al., 1984] utiliza los índices Gini para seleccionar la división máxima Gini.

***Bosques aleatorios (random forest)*:** son una combinación de árboles de decisión de tal manera que cada árbol depende de los valores de un vector aleatorio y con la misma distribución para todos los árboles en el bosque. Es decir, es un clasificador que consiste en una colección de clasificadores con estructura de árbol, $\{h(\mathbf{x}, \Theta_k), k = 1, ...\}$ donde Θ_k son vectores aleatorios independientes identicamente distribuidos y cada árbol emite un voto unitario para la clase más popular en la entrada $\mathbf{x}$ [Breiman, 2001].

Árboles extremadamente aleatorios: han sido introducidos por Pierre Geurts, Damien Ernst y Louis Wehenkel en [Geurts et al., 2006]. El algoritmo de crecimiento de árboles extremadamente aleatorios es similar al de

árboles aleatorios (*Random Trees*), pero hay dos diferencias:

- Los árboles extremadamente aleatorios no aplican el procedimiento de *bagging* para construir un conjunto de muestras de entrenamiento para cada árbol. El mismo conjunto de entrenamiento de entrada se utiliza para entrenar a todos los árboles.
- Los árboles extremadamente aleatorios escogen una división de nodos de manera muy extrema (tanto el índice variable como el valor de división de variables se escogen al azar), mientras que los bosques aleatorios encuentran la mejor división (óptima una por índice variable y valor de división de variables) entre un subconjunto aleatorio de variables.

2.4. Enfoques de clasificación de leucocitos

Algunos trabajos de investigación proponen clasificar los leucocitos en cinco clases: basófilo, eosinófilo, linfocito, monocito y neutrófilo, como en [Jiménez Díaz, 2007, Martínez Castro et al., 2014, Sarrafzadeh et al., 2014, Pang et al., 2015]. Otros realizan la clasificación de leucocitos por etapas, por ejemplo en [Zhao et al., 2017] inician clasificando leucocitos en basófilo, eosinófilo y otros leucocitos. Después de eso, clasifican aquellos que fueron asignados a la clase otros leucocitos en: linfocitos, monocitos y neutrófilos. Un caso similar se presenta en [Rezatofighi and Soltanian-Zadeh, 2011], donde se inicia la clasificación en basófilos y no basófilos. Posteriormente, aquellos que no son basófilos son clasificados en linfocitos, monocitos, eosinófilos y neutrófilos.

Capítulo 3

Desarrollo del proyecto

En este capítulo se presenta en la Sección 3.1 lo referente a las especificaciones del hardware y software. Posteriormente, en la Sección 3.2 se presentan y describen cada uno de los módulos desarrollados para este proyecto de tesis. Finalmente, en la Sección 3.3 se describe la arquitectura de la aplicación que integra los módulos de identificación y conteo de las células con el reconocimiento de leucocitos.

3.1. Especificaciones de hardware y software

La ejecución de los experimentos se realizó en una estación de trabajo (*workstation*) Dell Precision Tower 7910 cuyas características se describen en el Cuadro 3.1.

Modelo	Dell Precision Tower 7910
Memoria RAM	32 GB
Procesador	Intel Xeon(R) CPU E5-2620 v4 @ 2.10GHz x 32
Sistema Operativo	Ubuntu 16.04 LTS de 64 bits

Cuadro 3.1: Características del Equipo.

Los módulos desarrollados en este proyecto fueron implementados en el lenguaje de programación Python versión 3.6.4, con la ayuda de las siguientes bibliotecas:

- Sklearn versión 0.21.2

- Skimage versión 0.13.1
- Mahotas versión 1.4.4
- Numpy versión 1.14.0
- Pandas versión 0.22.0
- Scipy versión 1.0.0
- Opencv versión 3.4.3

3.2. Módulos del proyecto

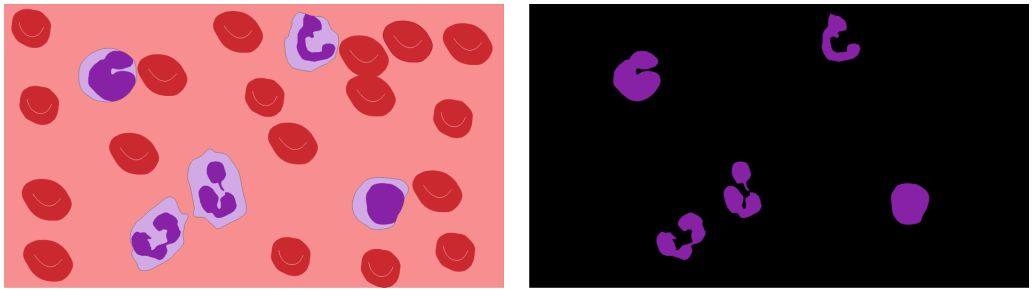
Esta sección tiene como finalidad describir cada uno de los módulos desarrollados, los cuales son derivados de la metodología presentada en la Sección 1.7. En la Subsección 3.2.1 se describe la primera etapa la cual se enfoca en aplicar diferentes técnicas de procesamiento digital de imágenes para identificar y contar automáticamente los leucocitos y eritrocitos. Finalmente, en la Subsección 3.2.2 se describen las fases para generar el mejor modelo para el reconocimiento de leucocitos.

3.2.1. Identificación y conteo de leucocitos y eritrocitos

Esta etapa consta de 4 módulos: identificación de núcleos, identificación de células, identificación de leucocitos y eritrocitos, y conteo de leucocitos y eritrocitos. Cada uno de estos módulos aplica técnicas de procesamiento digital de imágenes para lograr su objetivo. A continuación se describen cada uno de estos módulos.

Identificación de núcleos

El objetivo de este módulo es identificar los núcleos de leucocitos en imágenes de frotis sanguíneo. La Figura 3.1 ilustra el objetivo de este módulo, es decir, dada una imagen de frotis obtenemos como resultado los núcleos segmentados.



(a) (b)

Figura 3.1: Imagenen ilustrativa de un frotis sanguineo y su correspondiente segmentación de núcleos. a) Frotis sanguíneo. b) Núcleos segmentados.

La Figura 3.2 muestra el diagrama del proceso para lograr el objetivo de este módulo. La entrada es una imagen digital de frotis a la cual se aplica un preprocesamiento para reducir el ruido. Debido a que los núcleos determinan la posición de los leucocitos, se le aplica a la imagen anterior un algoritmo de segmentación que sea capaz de aislar los núcleos.

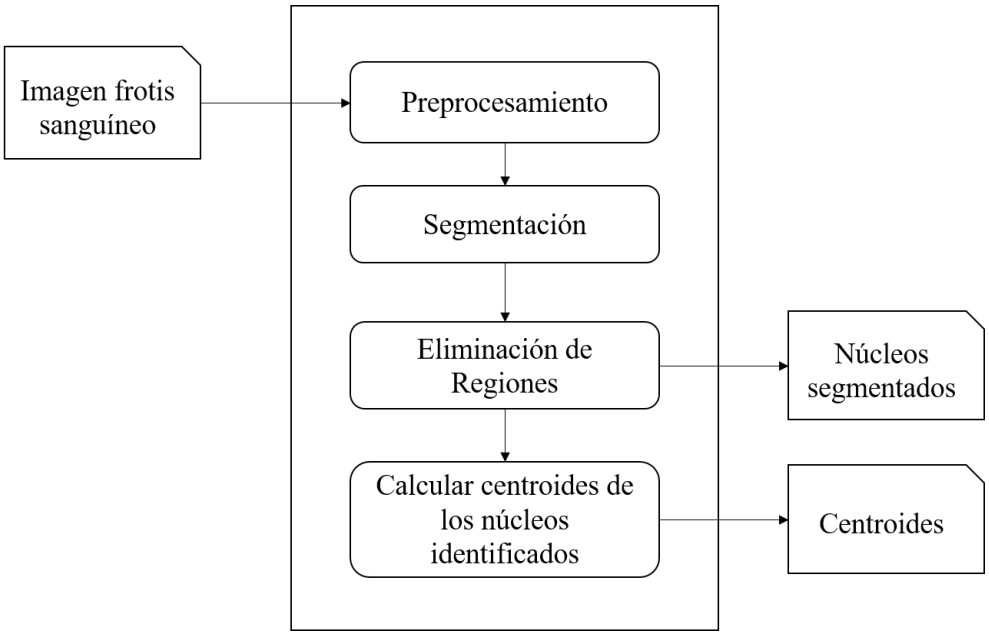


Figura 3.2: Diagrama del proceso para la identificación de núcleos.

Aunque la segmentación anterior aísla núcleos, de igual forma segmenta regiones erróneas. Esto se debe a que en el frotis hay elementos como el colorante precipitado, plaquetas u otros elementos que comparten valores de intensidades similares a los núcleos. En consecuencia, la fase de segmentación es deficiente.

Para reducir los errores derivados de una mala segmentación se crea una función que elimina, a través de un criterio de selección por área, aquellas regiones erróneas. Generando una identificación mejorada de los núcleos. Tras lograr la identificación de núcleos, la imagen $I_{nucleos}$ se etiqueta y calculan los centroides para cada una de las regiones. Los centroides se almacenan en una matriz $Centro_N$ que se define:

$$Centro_N = \begin{bmatrix} x_{cn1} & y_{cn1} & l_1 \\ . & . & \\ . & . & \\ x_{cnj} & y_{cnj} & l_j \end{bmatrix} \; j = 1 \ldots m \tag{3.1}$$

donde m es el número de núcleos, (x_{cnj}, y_{cnj}) son las coordenadas del centroide y l_j es la etiqueta del núcleo.

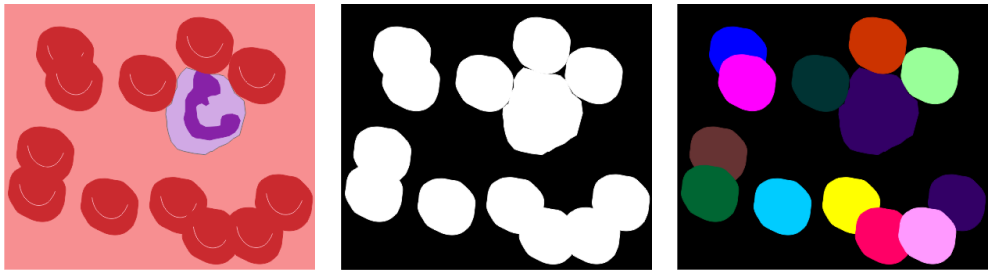
Para la identificación de núcleos se crearon las funciones que se muestran en el Cuadro 3.2.

Módulo	identificacion_nucleos_leucocito.py
Funciónes	identificacion_nucleos(...)
	elimina_Regiones(...)
	calcula_CentroidesNi(...)

Cuadro 3.2: Funciones principales del módulo identificación de núcleos.

Identificación de células sanguíneas

El módulo de identificación de células sanguíneas realiza el preprocesamiento para reducir el ruido y resaltar las regiones de interés (células). La imagen resultante del paso anterior se le aplica un algoritmo de segmentación para eliminar el fondo y dejar solo las células sanguíneas (leucocitos, eritrocitos o plaquetas).



(a) (b) (c)

Figura 3.3: Imágenes ilustrativas a) frotis sanguíneo, b) regiones fusionadas, c) segmentación ideal.

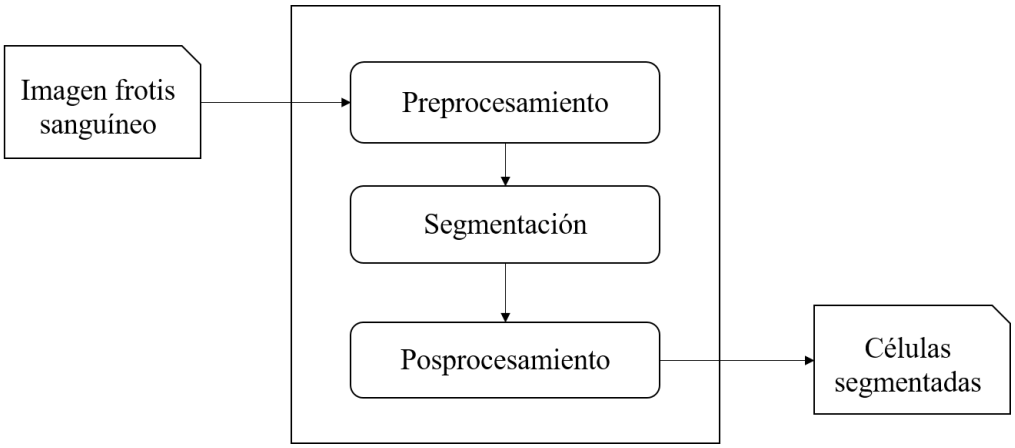


Figura 3.4: Diagrama del proceso para la identificación de células sanguíneas.

Algunas zonas del frotis carecen de condiciones óptimas por lo que genera problemas para tener buenos resultados en la segmentación. Las principales causas son células sobrepuestas una sobre la otra generando regiones segmentadas muy grandes que agrupan varios elementos, es decir, las regiones quedan fusionadas en una sola. Por esta razón, se aplica un postprocesamiento a la segmentación para separar las regiones fusionadas. La Figura 3.3 ilustra regiones fusionadas en una segmentación de frotis y una segmentación ideal de estas regiones fusionadas. En consecuencia, se minimizan errores y se previene que regiones fusionadas avancen a la siguiente fase. El proceso

utilizado para lograr lo mencionado anteriormente se muestra en la Figura 3.4.

Para la identificación de células sanguíneas se crearon las siguientes funciones que se muestran en el Cuadro 3.3.

Módulo	**identificacion_celulas.py**
Funciones	medianRGB(...)
	img_Segmentation(...)
	preprocesamiento_frotis(...)
	segmentacion_frotis(...)
	watershed_m2(...)
	identificacion_celulas_sanguineas

Cuadro 3.3: Funciones principales del módulo identificación de células sanguíneas.

Identificación de leucocitos y eritrocitos

El objetivo de este módulo es separar los leucocitos y eritrocitos. Para lograrlo se utilizan la imagen segmentada de células y la matriz de centroides de núcleos $Centro_N$ obtenidas en los dos módulos anteriores. La Figura 3.5 muestra el diagrama del proceso de éste módulo.

El proceso para separar los eritrocitos de los leucocitos inicia calculando el área de todas las regiones de la imagen y almacenándolas en un vector. Éste es ordenado de manera descendente para que las regiones más grandes queden al principio, esto es, porque los leucocitos tienen una mayor área en comparación con las demás células (ver Anexo A). El ordenamiento anterior garantiza que los leucocitos se posicionen al inicio del vector. A continuación, se realiza la separación de leucocitos con ayuda de una operación XOR y una función de apertura, esto se realiza hasta que el conteo de leucocitos llega a cero [Acharya and Kumar, 2018].

Cabe mencionar que el método anterior funciona bajo el supuesto de que se conoce previamente el número de leucocitos. Sin embargo, en este trabajo de tesis no conocemos dicha información. Por concecuencia, el proceso de separación de células rojas y blancas mencionado anteriormente es modificado para utilizar únicamente la información de los centroides de los núcleos $Centro_N$, y extraer simultáneamente los leucocitos. La Figura 3.5 muestra el

proceso para la extracción de leucocitos y eritrocitos en una imagen binaria. Primero, se aplica un algoritmo para etiquetar regiones en la imagen de células segmentadas. Posteriormente el área y la etiqueta de todas las regiones etiquetadas es calculada y almacenada en una matriz $areasGeneral$. Así la matriz se define:

$$areasGeneral = \begin{bmatrix} a_1 & l_1 \\ . & . \\ a_i & l_i \\ . & . \\ a_n & l_n \end{bmatrix} \tag{3.2}$$

donde n es el número de regiones que tiene la imagen segmentada de células, a_i es el área de una región y l_i es la etiqueta de la región.

A continuación, se ordena de manera descendente la matriz $areasGeneral$, así las regiones más grandes se posicionan al principio, es decir, garantizar que las regiones candidatas a leucocitos estén al principio y la búsqueda no sea exhaustiva.

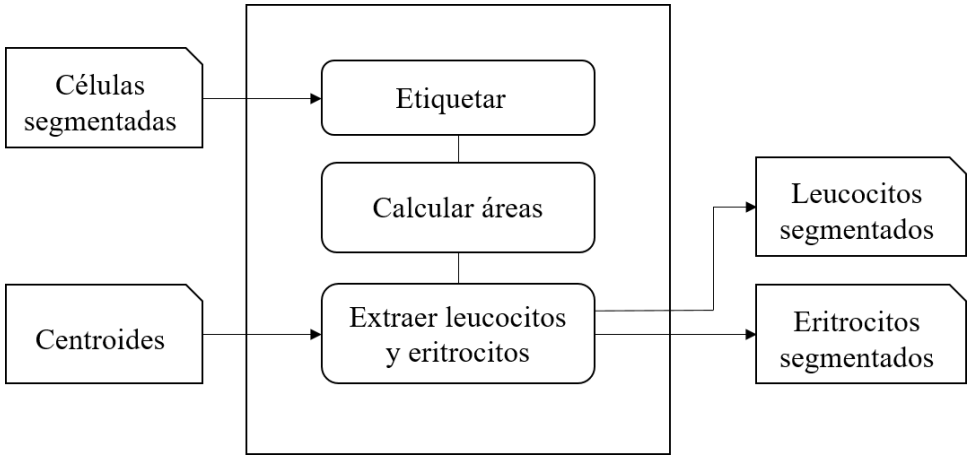


Figura 3.5: Diagrama del proceso para la identificación de leucocitos y eritrocitos.

Posteriormente, inicia la extracción de regiones candidatas a leucocitos, esto se realiza de manera iterativa, recorriendo la matriz de áreas y calculando los centroides (x_0, y_0) de la región R_j. A continuación, con la información anterior se verifica si R_j es un posible leucocito, esto es, si cumple con la siguiente condición:

$$(x_0 - \varepsilon_x \leq Centro_N[i,0] \leq x_0 + \varepsilon_x) \wedge (y_0 - \varepsilon_y \leq Centro_N[i,1] \leq y_0 + \varepsilon_y) \quad (3.3)$$

donde (x_0, y_0) son las coordenadas del centroide de R_j, $\varepsilon_x, \varepsilon_y$ son intervalos de desplazamiento que delimitan a la región y $Centro_N$ es la matriz de centroides de núcleos definida en la Ecuación (3.1).

La condición descrita anteriormente es de gran importancia en el proceso de identificación de regiones candidatas a leucocitos. Si la región cumple la condición se considera leucocito y se almacena en una imagen. Asimismo, se elimina esta región de la imagen de células segmentadas para dejar solo eritrocitos en esa imagen. La extracción de regiones candidatas a leucocitos se realiza en un ciclo y termina cuando ya se han detectado las regiones con núcleos, es decir, el tamaño de la matriz $Centro_N$ es igual a cero. El proceso anterior genera como resultado dos máscaras de segmentación donde la primera de ellas contiene solo leucocitos y la segunda solo eritrocitos.

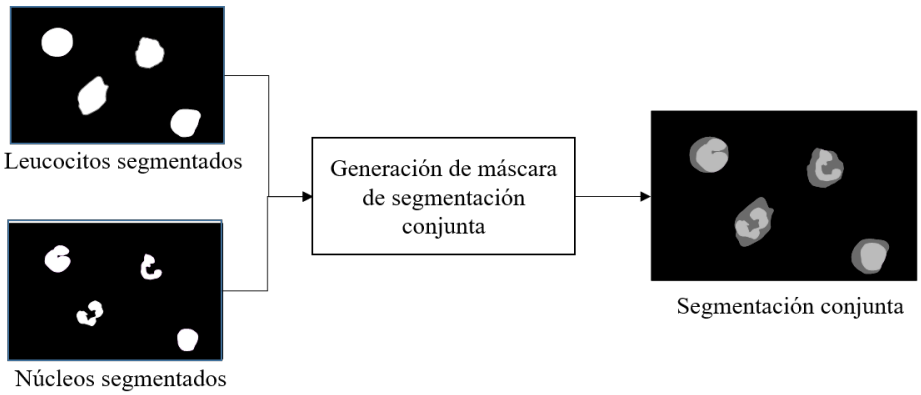


Figura 3.6: Proceso de creación de la máscara de segmentación conjunta.

Los procesos mostrados en la Figura 3.2 y la Figura 3.5 generan una máscara de segmentación de núcleos y una máscara de segmentación de leucocitos, respectivamente. Derivado de lo anterior, es posible generar una máscara de segmentación conjunta que muestre el núcleo y citoplasma del leucocito. En la Figura 3.6 se muestra el proceso para generar la máscara de segmentación conjunta, esto es, mediante la unión de núcleos y leucocitos identificados

y se asocia el nivel de intensidad en gris 170 con el núcleo y el 85 con el citoplasma.

El Cuadro 3.4 muestra las funciones del módulo de identificación de leucocitos y eritrocitos.

Módulo	**identificacion_WBC_RBC.py**
Funciones	identifica_WBC_RBC(...)
	calcula_CentroidesNi(...)
	calcula_CentroideRj_Ep(...)
	verifica_WBC(...)
	detect_Selected_Nucleo(...)
	arrayArea(...)

Cuadro 3.4: Módulo identificación de leucocitos y eritrocitos.

Conteo de leucocitos y eritrocitos

Debido a que se han identificado leucocitos y eritrocitos y se han generado máscaras de segmentación de estos, es posible realizar el conteo de eritrocitos y las regiones candidatas a leucocitos. La Figura 3.7 muestra el diagrama del proceso utilizado para realizar el conteo de las células.

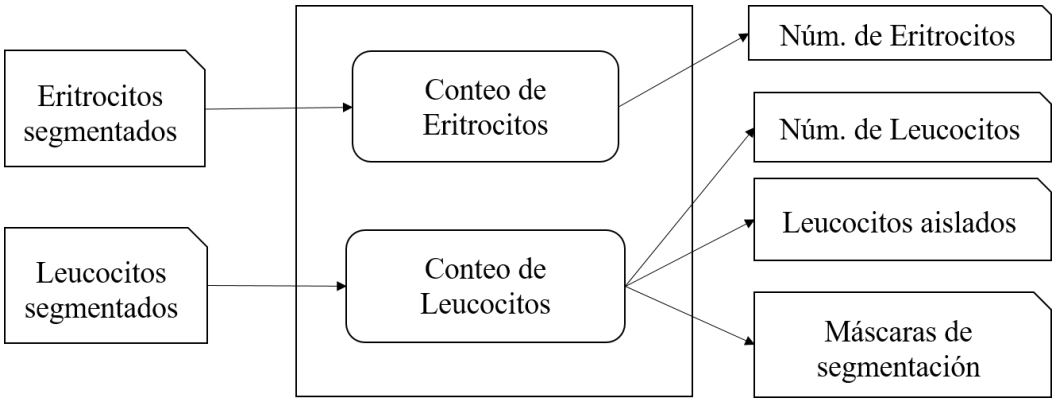


Figura 3.7: Diagrama del proceso para el conteo de leucocitos y eritrocitos.

El módulo está dividido en dos partes, la primera se enfoca en realizar el conteo de los eritrocitos identificados. Para lograr esto se toma como referencia el factor de forma propuesto en [Acharya and Kumar, 2018] y descrito en

la Subsección 2.2.4 . Asimismo, se agrega una restricción de área a la célula para evitar que pequeños elementos sean contabilizados. De esta manera, una región es contabilizada si cumple con estas dos restricciones planteadas. En la segunda fase se realiza el conteo de los leucocitos con un algoritmo de etiquetamiento.

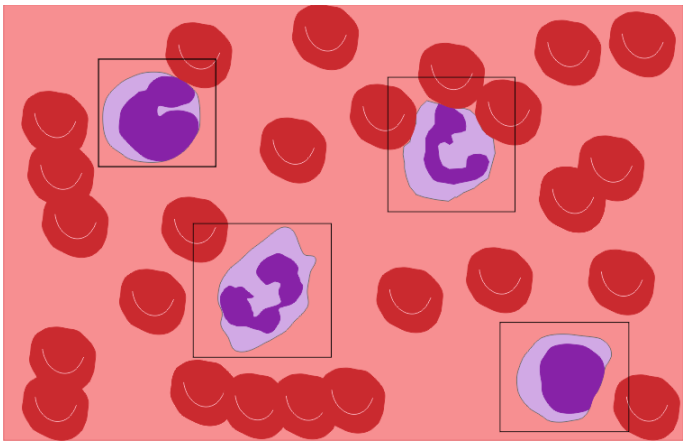


Figura 3.8: Candidatos a leucocitos identificados en un rectángulo.

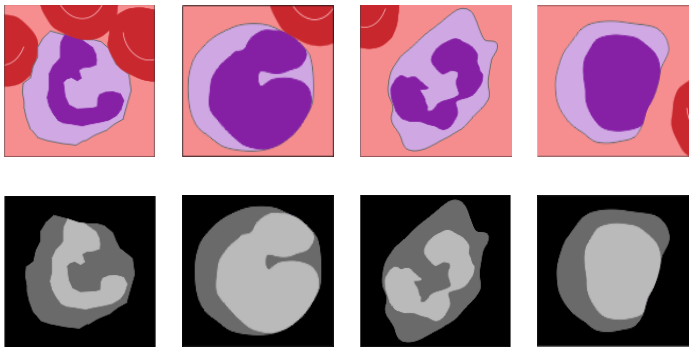


Figura 3.9: Imágenes de leucocitos aislados (primera fila) y máscaras de segmentación (segunda fila).

Una vez que los leucocitos han sido contados, es posible generar subimágenes que contienen completamente a la región (leucocito) con la finalidad de ser utilizadas para generar un conjunto de datos de leucocitos (ver Subsección 3.2.2). Esto se realiza con las coordenadas obtenidas de un rectángulo

que contiene completamente a la región (ver Figura 3.8). De esta manera un único leucocito queda aislado dentro del rectángulo y sus coordenadas sirven para recortar simultáneamente a los leucocitos y su máscara de segmentación. La Figura 3.9 muestra un ejemplo leucocitos aislados con sus respectivas máscaras de segmentación.

El Cuadro 3.5 muestra las funciones implementadas para el conteo.

Módulo	**conteo_Automatico.py**
Funciones	conteo_RBCs(..)
	arrayArea(..)
	conteo_WBCs(..)
	conteo_WBC_RBC(..)

Cuadro 3.5: Funciones principales del módulo de conteo.

3.2.2. Reconocimiento de leucocitos

Esta etapa consta de tres módulos (ver etapa 2 de la Figura 1.1): extracción de características, preprocesamiento de datos y clasificación de leucocitos. Cada uno de estos proveen información importante para el desarrollo de este proyecto de tesis. A continuación se describen cada uno de estos módulos.

Extracción de características

El proceso utilizado para realizar la extracción de características es mostrado en la Figura 3.10. Para realizar este proceso se requieren de dos fases, uno es la extracción de regiones de interés y la otra es el generador de características.

La fase de extracción de regiones de interés se enfoca en aislar cada uno de los componentes, es decir, el núcleo, el citoplasma y la célula. La extracción da como resultado tres tipos de imágenes. La primera es una imagen binaria que identifica núcleo, citoplasma y toda la célula. El segundo tipo muestra las regiones, pero con su equivalente en RGB, dicho de otra forma, muestra los niveles de intensidades en ese espacio de color, y el tercer tipo de imagen muestra los bordes del núcleo y citoplasma. La Figura 3.11 ilustra cada uno de los tipos de imágenes mencionadas anteriormente.

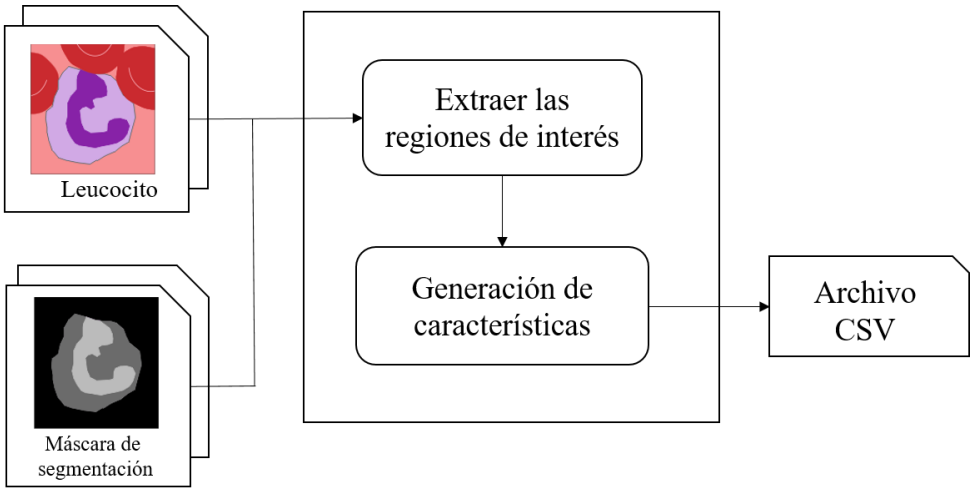


Figura 3.10: Diagrama del proceso para la generación de características.

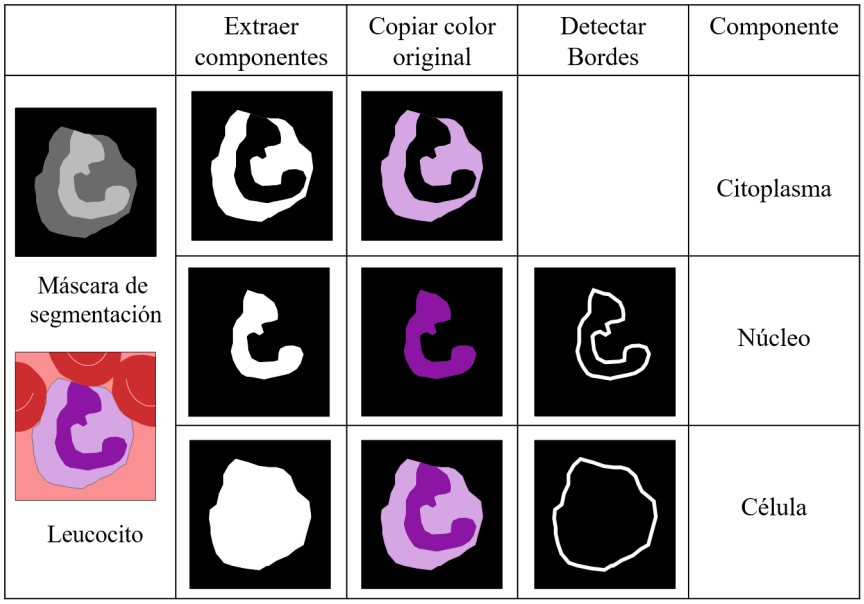


Figura 3.11: Imágenes de las regiones de interés del leucocito.

En la fase de generación de características se incluyen medidas que ya se han utilizado para el reconocimiento de leucocitos como son [Jiménez Díaz, 2007, Saraswat and Arya, 2014], área, perímetro, compacidad, dispersión, momentos invariantes de Hu del citoplasma [Sarrafzadeh et al., 2014], momentos

invariantes de Hu del núcleo [Pang et al., 2015], descriptores de Haralick [Sarrafzadeh et al., 2017, Shirazi et al., 2016]. También se usan los descriptores de color [Isaza et al., 2018]. Así como también los momentos de Zernike [Li et al., 2018]. El Cuadro 3.6 muestra las características calculadas a cada uno de los componentes de los leucocitos para generar el vector de características. Además, se indican cuales descriptores han sido calculados para cada componente de leucocito.

Descriptor	**Núcleo**	**Citoplasma**	**Célula**
Área	Si	Si	No
Perímetro	Si	Si	No
Compacidad	Si	Si	No
Dispersión	Si	No	Si
Momentos invariantes de Hu	Si	Si	Si
Haralick	Si	Si	Si
Momentos de Zernike	Si	Si	No
Media en cada uno de los espacios: RGB, HSV, YCbCr, L*a*b	Si	Si	Si
Varianza en cada uno de los espacios: RGB, HSV, YCbCr, L*a*b	Si	Si	Si

Cuadro 3.6: Lista de características utilizadas para generar el vector de características.

El proceso implementado en la fase de generación de características se describe a continuación. Primero, se convierten a escala de grises las imágenes de color del núcleo, citoplasma y célula para extraer los descriptores de Haralick y los momentos invariantes de Hu. Además, se calculan los momentos de Zernike al núcleo y citoplasma. A continuación, se calculan el área, perímetro, compacidad con las imágenes binarias del núcleo y citoplasma, con los bordes del núcleo y la célula se calcula la dispersión. Posteriormente, se realiza la conversión del espacio de color RGB a los espacios HSV, YCbCr y L*a*b. De esta manera se pueden calcular la media y la varianza en cada componente del espacio de color. Finalmente, el proceso de generación de características da un total de 193 características, es decir, 74 características para el núcleo, 73 características para citoplasma. y 46 características para la región que comprende toda la célula.

Preprocesamiento de datos

La fase de preprocesamiento de datos se describe a continuación. Primero, se genera de manera aleatoria y estratificada un conjunto de datos de entrenamiento y otro de prueba, éstos son de $M \times N$, donde N es el número de características y M es el número de ejemplos. Adicionalmente, se aplica a éstos una técnica de preprocesamiento de datos, esto es, normalizar los datos de entrenamiento y con base a los parámetros de normalización adquiridos, se normaliza el conjunto de prueba.

La normalización o estandarización de los datos es útil para evitar que las características con valores grandes tengan mayor influencia que las características con valores pequeños. Del conjunto de datos de entrenamiento y prueba, se ha de realizar una normalización. Ésta se realiza utilizando las biblioteca de *sklearn*, específicamente el módulo *sklearn.preprocessing*.

Reconocimiento de leucocitos

A partir de los datos normalizados se crean modelos para la clasificación de leucocitos, esto es, realizar la selección de características y realizar pruebas con algoritmos de clasificación para elegir el que proporcione el mejor rendimiento. Los módulos desarrollados para esta etapa se describen a continuación.

La selección de características es un proceso que elige automáticamente las características más relevantes mientras que elimina aquellas que, por ser redundantes o irrelevantes, afectan el desempeño del módulo. Por lo anterior, este módulo tiene el objetivo de reducir la dimensionalidad a través de la selección de características y obtener un subconjunto más pequeño de variables. En este contexto variable es sinónimo de característica.

La Figura 3.12 muestra el diagrama del proceso de selección de características. Cabe mencionar que este módulo fue inspirado en el trabajo de investigación en [Aushev et al., 2018], específicamente en algunos aspectos en la configuración experimental de sus pruebas realizadas.

Primero, se aplica un método de selección de características sobre el conjunto de entrenamiento, esto genera un subconjunto de k características. Entonces, se procede a seleccionar estas características para crear un nuevo conjunto de datos para entrenamiento y prueba con solo k características

(ver Figura 3.12a).

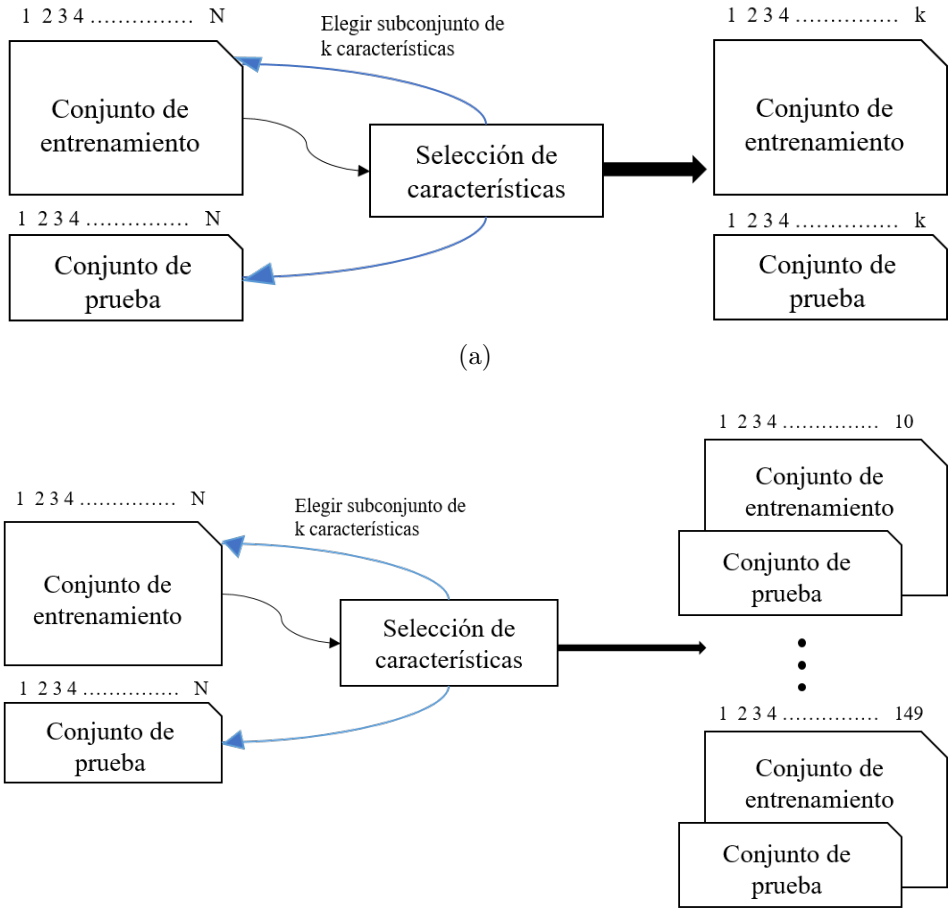


Figura 3.12: Diagrama proceso de la selección de características. a) Genera sólo un subconjunto de k características. b) Genera múltiples subconjuntos de k características.

Cabe mencionar que el proceso descrito anteriormente funciona cuando el método usado en la selección de características genera un subconjunto. Éste no considera aquellos métodos que requieren como parámetro el número k de características, esto sucede con los métodos filtro. Para abordar este problema, se ha propuesto una variación al proceso mostrado en la Figura 3.12a,

usando como referencia algunos aspectos en la configuración experimental presentada en [Aushev et al., 2018]. La Figura 3.12b muestra el diagrama del proceso de selección de características para este caso.

A continuación, se aplica un método de selección de características sobre el conjunto de entrenamiento, variando k en el rango $(10, 149)$. Entonces, se procede a seleccionar estas características para crear un nuevo conjunto de datos para entrenamiento y prueba con solo k características. Esto se realiza para cada valor k generando conjuntos con diferentes tamaños.

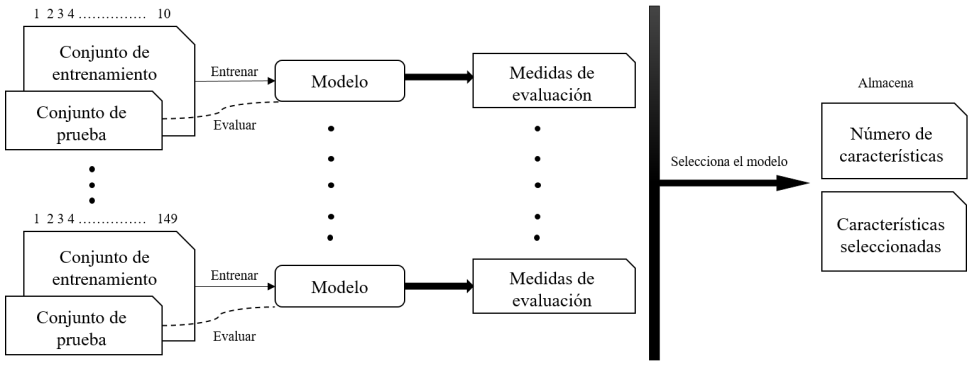


Figura 3.13: Diagrama proceso para la evaluación de cada subconjunto de características.

Una vez generados todos los conjuntos de datos se procede a realizar la evaluación de las k características para cada uno. Esto se realiza siguiendo el proceso mostrado en la Figura 3.13. Cada conjunto genera un modelo, éste es entrenado y probado siguiendo el proceso mostrado en la Figura 3.14, el cual es descrito en la siguiente subsección.

Módulo	**clasificador_stratificado.py**
Funciones	FS_A1(...)
	RFE_FS(...)
	FS_A6(...)
	indices(...)
	read_all(...)
	estratificado_typo(...)
	estratificado_subtypo(...)

Cuadro 3.7: Funciones principales del módulo de selección de características.

Finalmente, se selecciona el modelo que obtuvo el mejor rendimiento y por consecuencia almacenamos las características seleccionadas y el valor de k. Por lo anterior, se ha determinado el subconjunto de datos para el método con enfoque filtro con k características.

El Cuadro 3.7 muestra las funciones implementadas para realizar la selección de características.

Generación de modelos de reconocimiento de leucocitos

Los objetivos de este módulo son entrenar diversos algoritmos de clasificación y obtener los valores de los parámetros que maximicen el rendimiento en cada clasificador. Además, evaluar los modelos con los datos de prueba y así generar medidas de rendimiento. Cada modelo es entrenado con el subconjunto de k características generado por el módulo anterior. Para esto, se desarrolla el proceso mostrado en la Figura 3.14

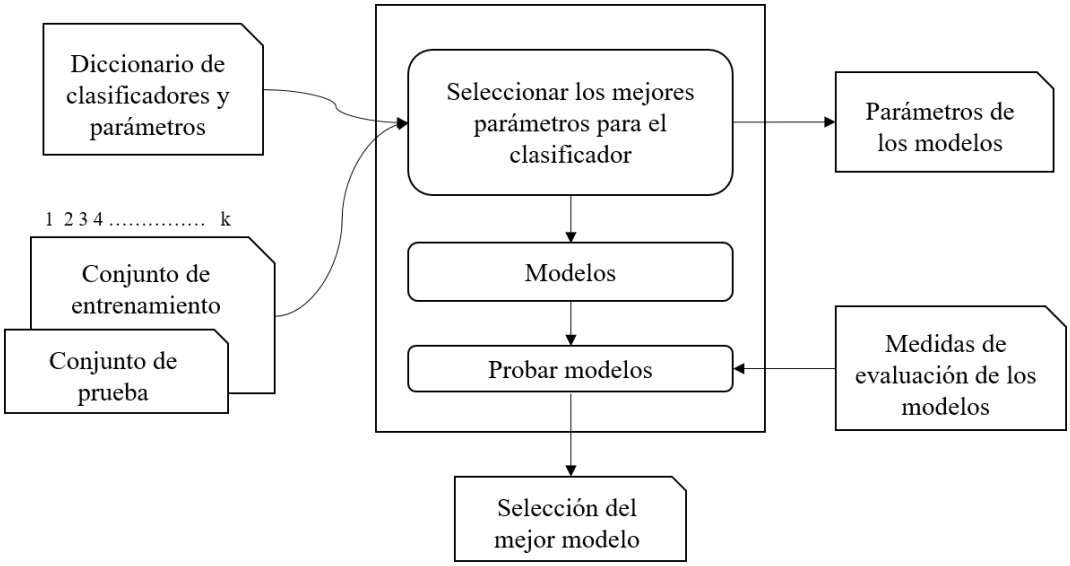


Figura 3.14: Diagrama del proceso para la generación de modelos.

En dicho módulo, se tienen diferentes configuraciones de entrada y son descritas a continuación:

- Diccionario de clasificadores y parámetros: Es el parámetro más importante debido a que éste contiene todos los clasificadores a utilizar. Cada

clasificador cuenta con sus parámetros y los valores de estos parámetros están definidos en intervalos. De esta manera, se requiere realizar una búsqueda en esos intervalos para obtener los mejores valores y generar un buen rendimiento en la etapa de entrenamiento del modelo.

- Datos de entrenamiento y prueba: Son los conjuntos de datos con k características seleccionada en el módulo anterior.

El proceso implementado en el módulo generación de modelos se describe a continuación. Primero, se entrena un algoritmo del diccionario de entrada. Es decir, se seleccionan los mejores valores de los parámetros del clasificador, buscando los valores de parámetros que elevan el rendimiento de clasificación en la etapa de entrenamiento del modelo. Tras generar el modelo con el mejor rendimiento en clasificación y obtener los mejores valores de parámetros, éste es evaluado con los datos de prueba y se calculan las medidas de evaluación. Los pasos mencionados anteriormente se realizan para cada uno de los clasificadores que están en el diccionario. Derivado de lo anterior, el conjunto de datos con k características generan L modelos con sus respectivos parámetros.

El Cuadro 3.8 muestra las funciones implementadas para la generación de los modelos.

Módulo	**init.py**
Funciones	all_Dicts(...)
	main(...)
	batch_classify(..)
	BER(...)

Cuadro 3.8: Funciones principales del módulo de generación de modelos.

3.3. Aplicación de muestra

En las secciones anteriores se han mostrado los módulos correspondientes para la identificación y conteo de leucocitos y eritrocitos, así como el reconocimiento de los cinco principales leucocitos. En esta sección se presenta el desarrollo de una aplicación que integra ambas actividades y muestra el uso práctico de la metodología mostrada en la Figura 1.1. La aplicación se divide en dos partes importantes y para lograrlo se utilizó el *framework* Angular

debido a que éste tiene el propósito de realizar la separación entre el *frontend* y el *backend* en una aplicación web. Además, permite el desarrollo de una aplicación de página única (SPA, por sus siglas en inglés).

Una aplicación SPA creada con Angular es de una página única, en la cual la navegación entre secciones y páginas de la aplicación, así como la carga de datos, se realiza de manera dinámica, casi instantánea. De manera asíncrona se hacen peticiones al servidor y sobre todo sin refrescar la página en ningún momento. Por esta razón se utilizó Angular para el desarrollo de la aplicación Web.

Con respecto al *backend* de la aplicación, aquí se ejecutan los módulos que se encargan de identificar y contar las células, así como reconocer los leucocitos. Mientras por el lado del *frontend* se realizan peticiones al servidor a través de la aplicación Web. La arquitectura del sistema es mostrada en la Figura 3.15.

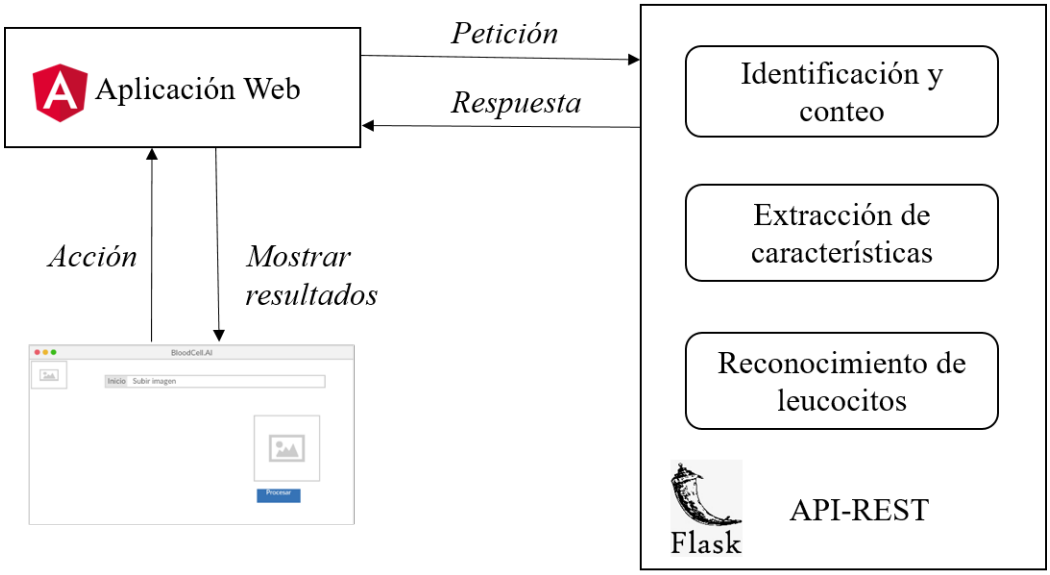


Figura 3.15: Arquitectura de la aplicación desarrollada.

El flujo de la aplicación es el siguiente. Primero el usuario entra a la aplicación Web desde su navegador y realiza una petición, por ejemplo, procesar una imagen de frotis. A continuación, la aplicación realiza una petición HTTP a la API REST. Segundo, la API al recibir la petición la procesa

y devuelve como respuesta un objeto del tipo JSON a la aplicación Web. Posteriormente, esta respuesta es procesada y el usuario lo visualiza en la interfaz.

El usuario navega entre secciones y páginas de la interfaz diseñada. Esto es, subir imágenes de frotis sanguíneo y visualizar resultados de conteo o reconocimiento. En el Cuadro 3.9 se describen los principales secciones de la aplicación Web.

URL	**Descripción**
/ /home	Muestra la página principal.
/subir-frotis	Muestra la interfaz que permite subir una imagen de frotis sanguíneo.
/show-frotis/:id/:CE/:CL	Muestra una pantalla con los resultados de conteo de leucocitos y eritocitos.
/recognition/:id/:num	Muestra una pantalla con los resultados de reconocimiento de leucocitos.

Cuadro 3.9: Rutas más importantes para la navegación en la interfaz.

Toda la lógica para realizar la identificación y conteo de células, además el reconocimiento de leucocitos se encuentra desarrollada en el *backend* de la aplicación. Éste fue desarrollado con ayuda del *framework Flask*. Los endpoints (rutas de conexión que responden a una petición) de la API REST desarrollada se describen en el Cuadro 3.10. Los componentes más importantes del *backend* son los siguientes:

Identificación y conteo: Es un paquete que se encuentra alojado en la parte del servidor y está conformado por los cuatro módulos descritos en la Subsección 3.2.1. En la Figura 1.1 se observa la secuencia de pasos para realizar la identificación y conteo, específicamente la etapa uno.

Extracción de características: Este paquete se encarga de generar las características de la imagen de leucocito, siguiendo el proceso mostrado en la Figura 3.10.

Reconocimiento de leucocitos: El modelo para realizar el reconocimiento de leucocitos se encuentra alojado en el servidor. En la Figura 3.16 se observa la secuencia para realizar la clasificación de leucocitos.

HTTP	URL	Descripción
POST	/upload-file	Recibe una imagen de frotis, la procesa y devuelve como respuesta un JSON con los resultados de conteo de leucocitos y eritrocitos, y el id de la imagen.
GET	/upload/<filename>	Muestra una imagen desde el servidor dado el nombre del mismo.
GET	/FS/<id_img>/<num>	Recibe el id del frotis y el número de leucocitos identificados y realiza el reconocimiento de los n leucocitos. Devuelve como respuesta un JSON con el id de la imágen y los resultados de reconocimiento

Cuadro 3.10: End-point más importantes de la API-REST.

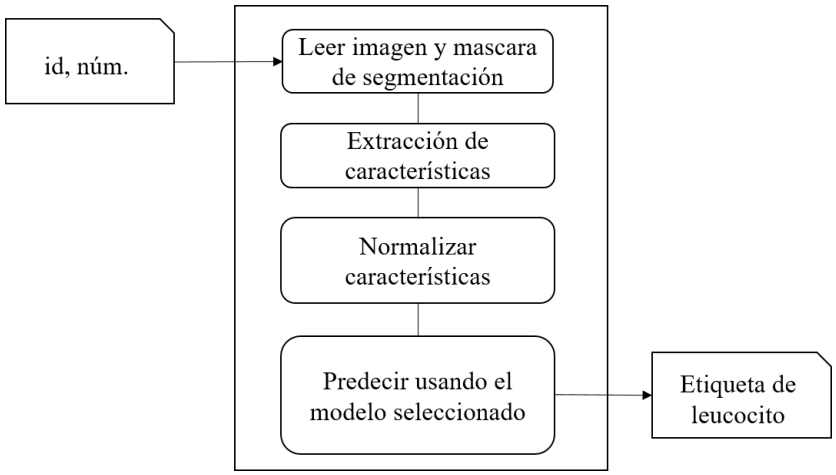


Figura 3.16: Diagrama del proceso para el reconocimiento de leucocitos.

Para mayor información sobre el uso de la aplicación Web, ver Anexo C.

Capítulo 4

Resultados

En este capítulo se presentan, en la primera sección, la descripción del conjunto de imágenes de frotis sanguíneo utilizado para llevar a cabo la fase experimental de este proyecto de tesis. En la segunda sección se presentan los resultados de identificación de núcleos, identificación de células sanguíneas y la identificación y conteo de eritrocitos, y regiones candidatas a leucocitos. Aquí se muestran los resultados de esta etapa de la metodología utilizando 3 imágenes con diferente calidad. En la tercera sección se presentan los enfoques de clasificación utilizados para clasificar los leucocitos. Además, se explica cómo se generó el conjunto de datos a utilizar. En la cuarta sección se desarrollan los experimentos realizados para la selección de un algoritmo de aprendizaje computacional adecuado para el problema de reconocimiento de leucocitos. También, se explican los experimentos realizados para seleccionar el mejor subconjunto de características. Finalmente, se muestran los resultados de reconocimiento de leucocitos con los mejores modelos y el mejor subconjunto de características.

4.1. Conjunto de datos

Los conjuntos de imágenes microscópicas de frotis sanguíneo empleados en este proyecto de tesis se tomaron de dos bases de datos públicas. En primer lugar, la base de datos de imágenes de leucemia linfoblástica aguda (Acute Lymphoblastic Leukaemia image database, ALL-IDB)[Labati et al., 2011] proporciona un total de 108 imágenes de tamaño 2592×1944 píxeles. Esta base contiene 59 imágenes de pacientes normales y 49 imágenes de pacientes

con leucemia linfoblástica aguda.

Este proyecto de tesis se enfoca en el reconocimiento de los cinco tipos de leucocitos sin presentar leucemia u otro tipo de célula anormal en la sangre periférica. Por esta razón, sólo se utilizan las imágenes derivadas de pacientes sanos. De dicha base, el experto registró un total de 102 leucocitos identificados para las 59 imágenes. El Cuadro 4.1 muestra la distribución de leucocitos para este conjunto de imágenes llamado DS_1.

Núm. Imágenes	**Basófilo**	**Eosinófilo**	**Linfocito**	**Monocito**	**Neutrófilo**	**Total**
59	1	2	67	6	26	102

Cuadro 4.1: Distribución de leucocitos del conjunto DS_1 respecto a una clasificación de cinco clases.

El segundo conjunto de imágenes utilizado, fue proporcionado por la Universidad de Ciencias Médicas de Isfahan [Sarrafzadeh et al., 2014] y está disponible en la página oficial de MISP [Center, 2019]. La base de datos contiene 100 imágenes de tamaño 3871 x 2592 píxeles adquiridas de frotis sanguíneo de diez pacientes. De estas imágenes se registran 117 leucocitos en 90 imágenes. Las 10 imágenes sobrantes se descartaron debido a que el experto no pudo determinar el tipo de leucocito en la imagen debido a las malas condiciones de estas. El Cuadro 4.2 muestra la distribución de leucocitos para este conjunto de imágenes denominado DS_2.

Núm. Imágenes	**Basófilo**	**Eosinófilo**	**Linfocito**	**Monocito**	**Neutrófilo**	**Total**
90	4	5	28	10	70	117

Cuadro 4.2: Distribución de leucocitos del conjunto DS_2 respecto a una clasificación de cinco clases.

Para generar un modelo de clasificación de leucocitos es necesario contar con muestras suficientes de cada una de las células. Por esta razón, se unen los conjuntos DS_1 y DS_2 para mostrar en el Cuadro 4.3 que el porcentaje de basófilos, eosinófilos y monocitos es aún muy pequeño en comparación con los neutrófilos y linfocitos. Por lo anterior, es necesario sumar a las clases minoritarias más ejemplos de imágenes. Para lograr esto, se

han agregado imágenes de dos conjuntos de imágenes. Los nuevos conjuntos de imágenes fueron obtenidos del trabajo de [Jiménez Díaz, 2007] y de [Zheng et al., 2018]. Estos conjuntos cuentan con imágenes de leucocitos, y su respectiva máscara que segmenta el núcleo y citoplasma de la célula. Para el caso de [Zheng et al., 2018], estas imágenes fueron adquiridas de CellaVision®. A este nuevo conjunto de imágenes de leucocitos se le denomina $Union_A$. En el Cuadro 4.4 se muestra el resultado de hacer la agregación de los conjuntos de datos mencionados anteriormente con DS_1 y DS_2.

Leucocito	**Etiqueta**	**Núm. leucocitos**	**Porcentaje**
Basófilo	C1	5	2.283 %
Eosinófilo	C2	7	3.196 %
Linfocito	C3	95	43.379 %
Monocito	C4	16	7.306 %
Neutrófilo	C5	96	43.836 %
	Total	219	100 %

Cuadro 4.3: Distribución de leucocitos de la unión de DS_1 y DS_2.

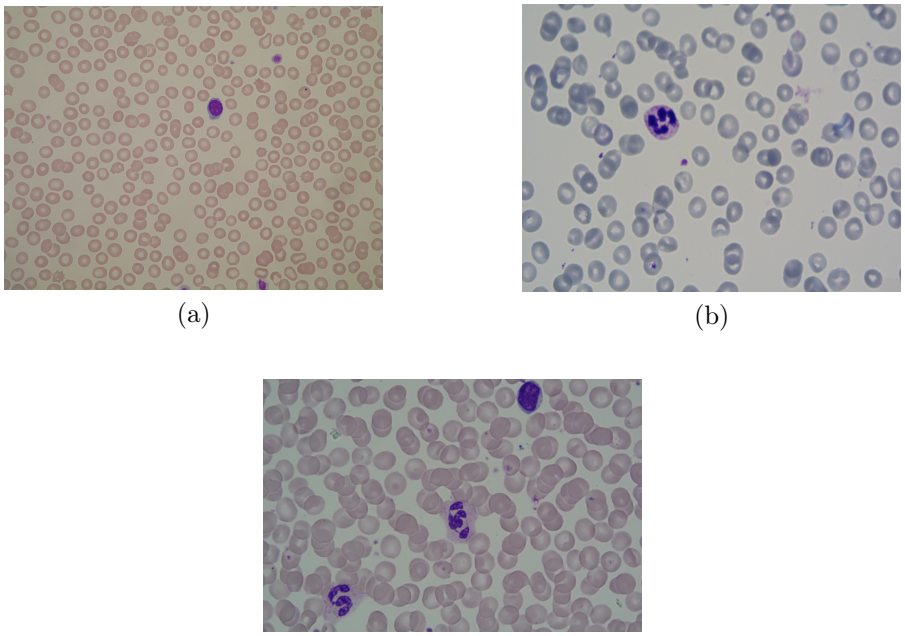
Leucocito	**Etiqueta**	**Conjunto inicial**	**Lucio**	**Cella Vision**	**Total**	**Porcentaje**
Basófilo	C1	5	8	3	16	5.047 %
Eosinófilo	C2	7	13	9	29	9.148 %
Linfocito	C3	95	0	0	95	29.968 %
Monocito	C4	16	49	16	81	25.552 %
Neutrófilo	C5	96	0	0	96	30.284 %
	Total	219	70	28	317	100 %

Cuadro 4.4: Distribución de leucocitos para $Union_A$.

4.2. Resultados de identificación y conteo

La metodología descrita en la Figura 1.1 y los cuatro módulos presentados en la Subsección 3.2.1 son utilizados para lograr la identificación y conteo de leucocitos y eritrocitos presentes en una imagen microscópica de frotis sanguíneo. Estos cuatro módulos forman parte de la aplicación planteada en el Capítulo 3.

La primera etapa tiene como objetivo identificar y contar automáticamente los leucocitos y eritrocitos presentes en una imagen microscópica de frotis sanguíneo. Esta etapa debe realizarse en tres fases, la primera fase tiene como objetivo identificar los núcleos de leucocitos y calcular los centroides de estos. La segunda fase se enfoca en identificar todas las células sanguíneas existentes en el frotis sanguíneo. La tercera fase se dedica a separar las células previamente identificadas en leucocitos y eritrocitos, y la cuarta fase se enfoca en contar automáticamente las células identificadas, además de generar subimágenes de los leucocitos.



(a) (b)

(c)

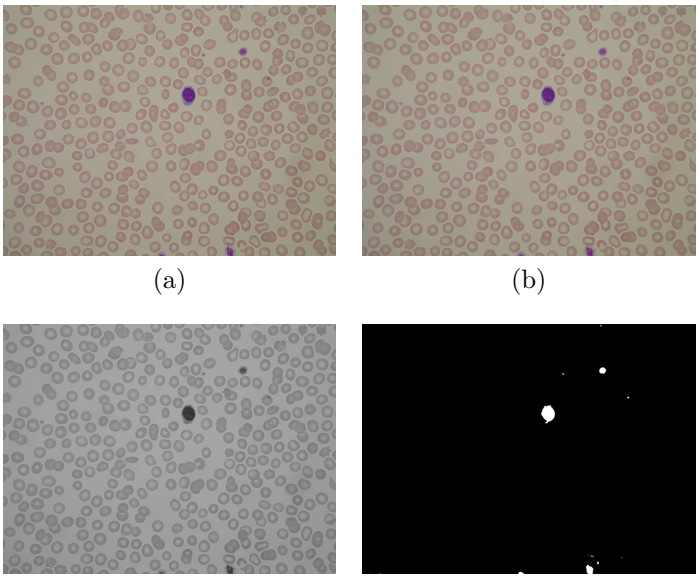
Figura 4.1: Imagen de frotis sanguíneo de calidad a) alta, b) media y c) baja calidad.

Para presentar los resultados de esta etapa el especialista consultado plantea dos criterios para seleccionar una imagen de alta, media y baja calidad. El primer criterio es sobre la calidad de tinción del frotis, esto es con el fin de evitar coloraciones defectuosas. El segundo criterio es sobre la zona de lectura ideal, es decir, en esa zona las células aparecen separadas lo suficiente para reconocerlas [Jaime Pérez and Gómez Almaguer, 2005].

Por los criterios anteriores, el especialista seleccionó tres imágenes. La primera cumple con todos los criterios por lo que se considera de calidad alta. La segunda visualmente cumple el criterio de zona de lectura, pero tiene una distribución de eritrocitos medianamente aglutinados comparada con una imagen de alta calidad, por lo que se considera de calidad media. La tercera imagen no cumple con ninguno de los criterios por lo que es de mala calidad. La Figura 4.1 muestra las imágenes microscópicas del frotis sanguíneo seleccionadas con los tres tipos de calidad.

4.2.1. Identificación de núcleos de leucocitos

En esta parte se presentan los resultados obtenidos en la fase de identificación de núcleos de leucocitos, aplicados a las tres imágenes de frotis sanguíneo seleccionadas anteriormente y utilizando el módulo descrito en la Subsección 3.2.1. La entrada de este módulo es una imagen microscópica de frotis en el espacio de color RGB.



(a) (b) (c) (d)

Figura 4.2: Ejemplo del proceso del preprocesamiento y segmentación. a) Imagen de alta calidad. b) Filtro mediana. b) Conversión a escala de grises. d) Umbralización de Yen.

Primero, en la etapa de preprocesamiento un filtro de mediana con elemento estructurante en forma de disco de $radio = 5$, es aplicado a cada canal de la imagen para reducir el ruido. Además, la imagen se convierte a escala de grises para eliminar la saturación y matiz, y conservar la luminosidad. Esta conversión se realiza con la Ecuación (2.1). A continuación, se busca segmentar los núcleos de los leucocitos debido a que los núcleos de los leucocitos son regiones de bajo contraste. Por lo anterior, el método de umbralización de Yen's es utilizado. Un ejemplo de este proceso es mostrado en la Figura 4.2.

Aunque la segmentación anterior aísla núcleos, de igual forma segmenta regiones erróneas, esto se debe a que los valores de intensidades de otros componentes sanguíneos como las plaquetas son similares a los niveles de intensidad en núcleos de los leucocitos. Con el objetivo de eliminar aquellas regiones que no corresponden a núcleos, un criterio de selección por áreas es aplicado. El proceso se describe a continuación.

1. La operación morfológica de apertura (*opening*) es aplicado a la imagen I_{mask}. Con esta operación, la imagen I_{open} contiene una menor cantidad del ruido tipo sal.

2. La imagen I_{open} es etiquetada y se calcula y almacena el área de todas las regiones en el vector A.

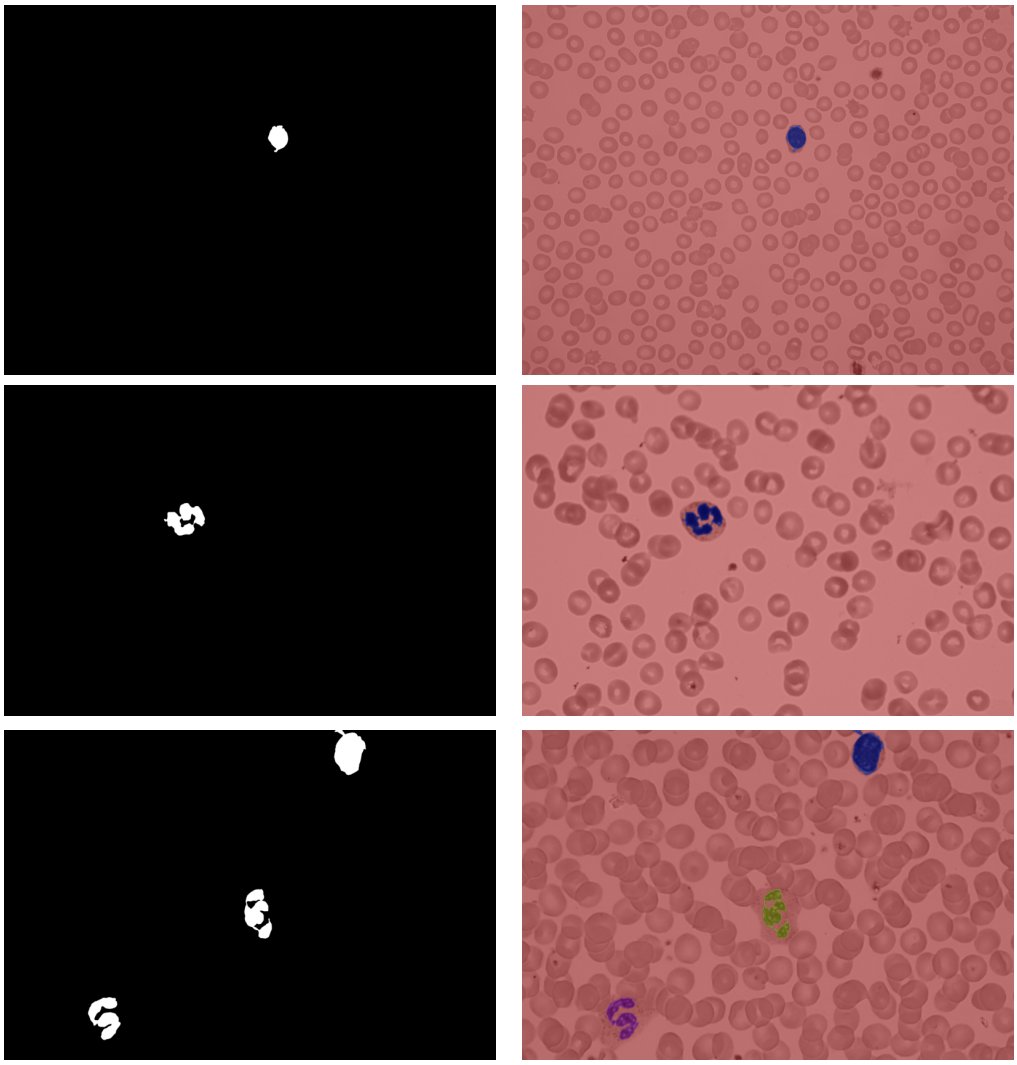
3. La media μ y desviación estándar σ de A es calculada.

4. Las regiones en la imagen I_{open} que son mucho más grandes que las demás se asume que son núcleos y las demás regiones deben eliminarse. Para hacer esto, aquellas regiones que cumplen con la siguiente condición $|A - \mu| <= \sigma$ se consideran regiones ruido y se eliminan. Con esta operación, la imagen $I_{nucleos}$ debe contener los núcleos de leucocitos.

Finalmente, se etiquetan las regiones de $I_{nucleos}$ para generar la imagen I_{label}. Se calculan los centroides de cada región etiquetada (Ecuación (3.1)) y se almacenan en $Centro_N$. Realizar este paso es importante, debido que los centroides determinan la ubicación de los leucocitos.

El procedimiento anterior genera como resultado una identificación mejorada de núcleos (ver Figura 4.3). En la primera columna se muestran los

núcleos segmentados de los leucocitos. La segunda columna muestra los núcleos identificados sobre la imagen original para una mejor visualización de éstos.



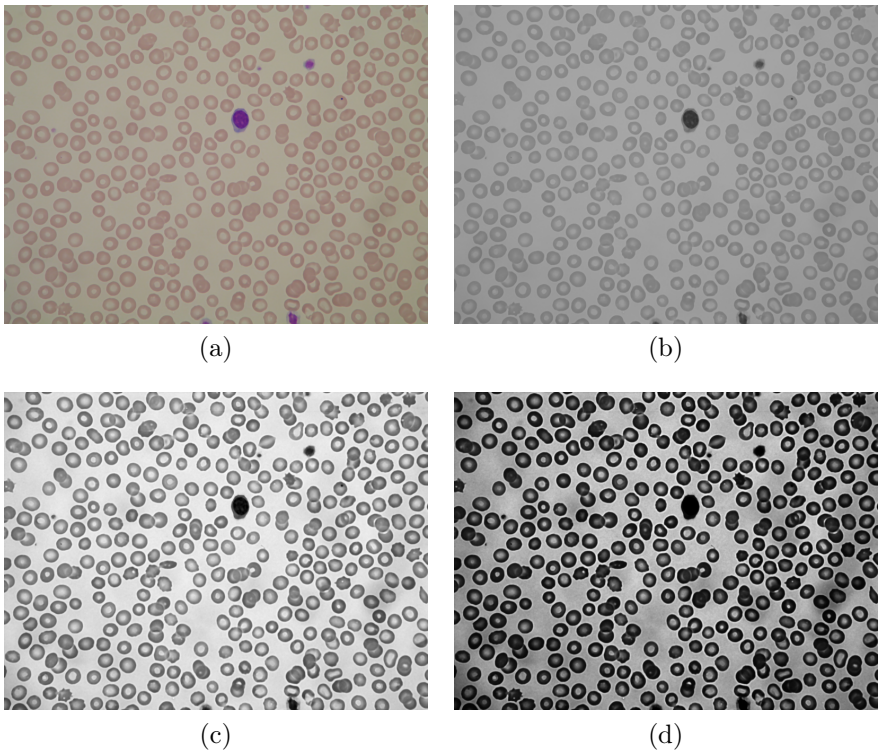
(a) (b)

Figura 4.3: Resultados de la identificación de núcleos en imágenes, en orden descendente por renglón: alta, media y baja calidad. a) Núcleos segmentados. b) Núcleos superpuestos a la imagen original.

En resumen, la calidad de las imágenes de frotis sanguíneo no afecta la identificación de núcleos, esto es porque lo núcleos de los leucocitos siempre son visibles independientemente de la zona de lectura del frotis.

4.2.2. Identificación de células sanguíneas

En esta subsección se presentan los resultados de identificación de células en las imágenes seleccionadas en la Sección 4.1. Realizar una identificación de células implica utilizar múltiples técnicas de procesamiento digital de imágenes y para lograrlo se aplica el procedimiento descrito en la Subsección 3.2.1. Este proceso se desarrolla en tres pasos, prepreprocesamiento, segmentación y postprocesamiento. Cada uno de los métodos y parámetros aplicados en cada paso fueron elegidos de manera experimental.



(a) (b) (c) (d)

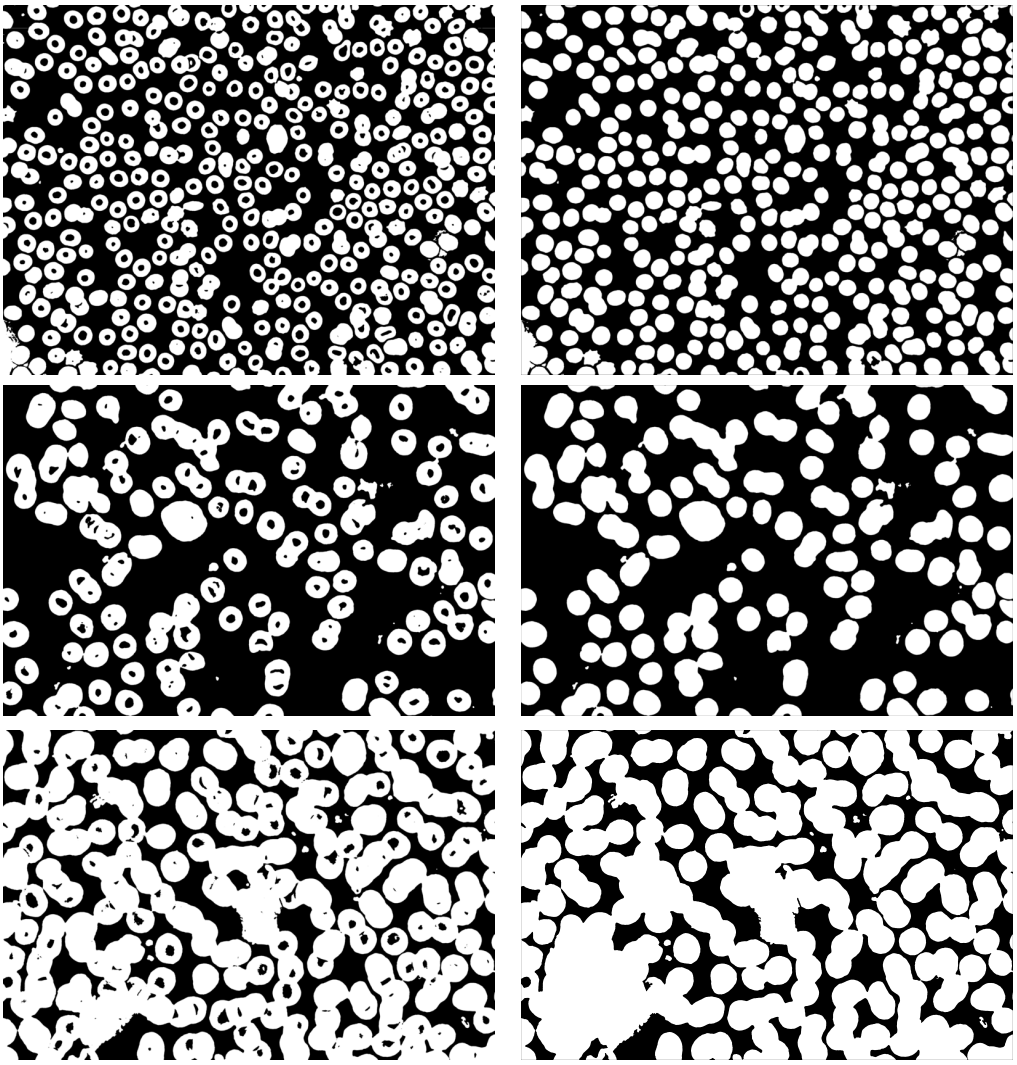
Figura 4.4: Ejemplo del preprocesamiento para la imagen de alta calidad. a) Filtro mediana. b) Conversión a escala de grises. c) *CLAHE*. d) Transformación gamma.

Con el fin de mejorar la imagen se emplean diversas técnicas de preprocesamiento. Primero, un filtro mediana con elemento estructurante en forma de disco de $radio = 5$ es aplicado a cada canal de la imagen para reducir el ruido. Después se convierte a escala de grises para eliminar la saturación y matiz en la imagen y conservar la luminosidad. Esta conversión se realiza con la Ecuación (2.1).

A continuación se utiliza el algoritmo de ecualización adaptativa del histograma limitada en contraste (*CLAHE*, por sus siglas en inglés) para mejorar el contraste local, es decir, los detalles locales se pueden mejorar en regiones más obscuras o más claras. Debido a que el fondo de la imagen es una zona muy homogénea se requiere utilizar el parámetro $clip - limit$ para limitar el realce de contraste en estas zonas. Por lo anterior, se ha establecido el valor de $clip - limit = 0.01$.

Debido a que el contraste de la imagen anterior es alto, adicionalmente se realiza un ajuste a las intensidades para oscurecer la imagen y garantizar una buena segmentación. Por esta razón, se aplica una transformación gamma (con $\gamma = 3$). La Figura 4.4 muestra el proceso de preprocesamiento. La primera fila muestra los resultados de usar un filtro mediana y convertir a escala de grises. La segunda fila muestra los resultados de aplicar *CLAHE* sobre la imagen en escala de grises y finalmente se realiza la transformación gamma a la imagen.

La imagen resultante del paso anterior se convierte a su forma binaria usando la umbralización global de Otsu. Como se puede observar en la Figura 4.5.a), algunas células en las imágenes no logran alcanzar el umbral causando huecos en las regiones que generalmente son eritrocitos. Una forma de resolver el problema es aplicarle a la imagen segmentada una operación morfológica de relleno de hoyos. Adicionalmente, se aplica una operación de erosión con elemento estructurante en forma de disco de $radio = 1$ para eliminar regiones muy pequeñas. Los resultados son almacenados en la imagen I_{seg}. La Figura 4.5.b) muestra los resultados al utilizar estas operaciones morfológicas sobre las imágenes segmentada de células.



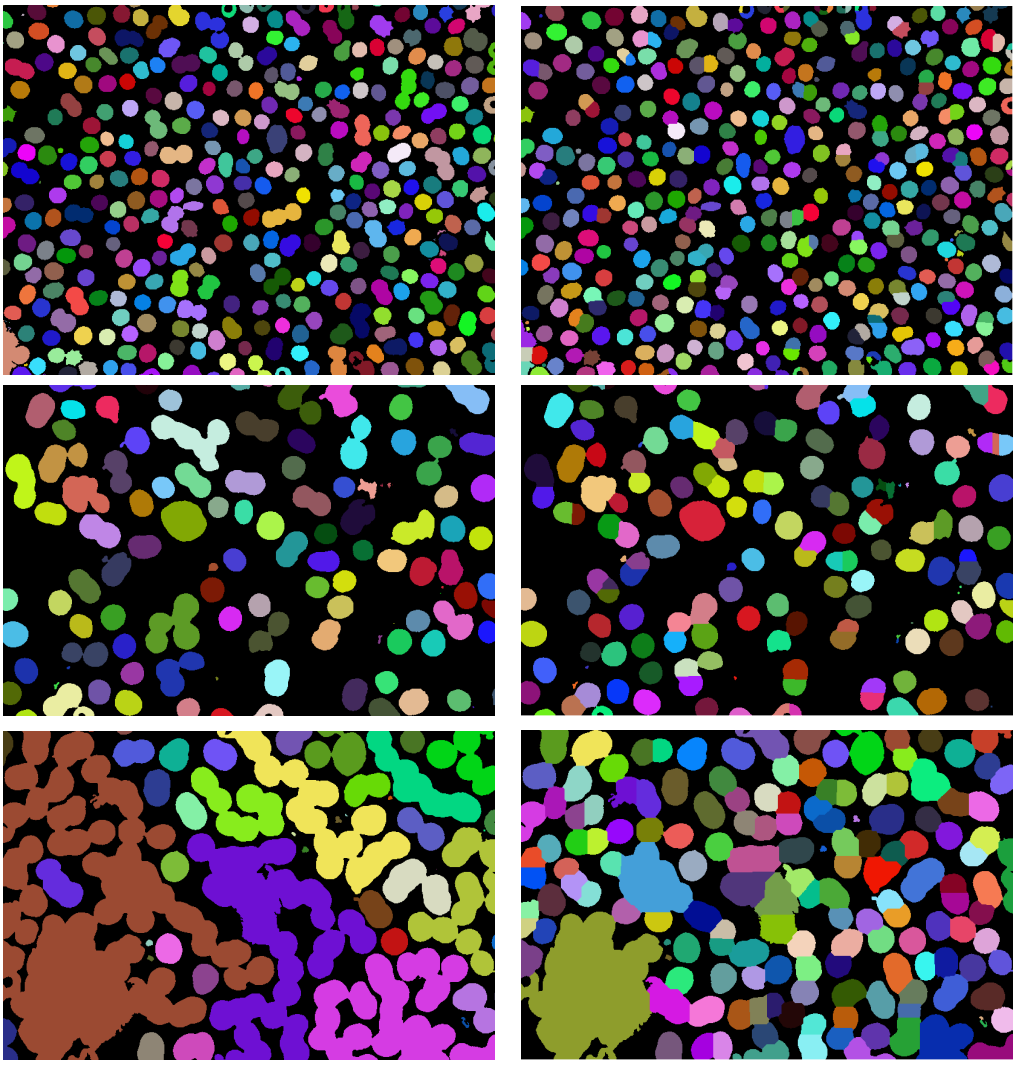
(a) (b)

Figura 4.5: Resultados de segmentación de células en imágenes, en orden descendente por renglón: alta, media y baja calidad. a) Umbralización global de Otsu. b) I_{seg}.

Respecto a los resultados de segmentación presentados anteriormente, se hace la siguiente observación. Algunas células superpuestas al ser segmentadas se fusionan en una sola región generando errores (ver Figura 3.3). Por lo anterior, las condiciones en la calidad del frotis sanguíneo influyen directamente en los resultados de segmentación. Entonces es necesario aplicar un postprocesamiento para separar las regiones fusionadas y minimizar estos errores de segmentación.

Se utiliza como método de postprocesamiento a la transformación *watershed* controlada por marcadores para resolver la problemática anterior. En el caso de la segmentación de células sanguíneas los marcadores internos a utilizar en *watershed* se definen como los máximos locales de la matriz de distancias [Miao and Xiao, 2018]. A estos marcadores encontrados en la matriz de distancias por convención se aplica una dilatación para ayudar a eliminar máximos locales triviales y así obtener una elección óptima de los marcadores. Finalmente, los marcadores y la imagen segmentada se pasan como parámetros a la transformada de *watershed* para tener una segmentación mejorada de las regiones que han quedado fusionadas.

Al aplicar el postprocesamiento, se logra separar la mayoría de las regiones fusionadas, pero en otras regiones muy sobrepuestas (baja calidad) no es posible. En la Figura 4.6 se muestran los resultados del postprocesamiento para los tres tipos de calidad de un frotis e ilustra que las regiones separadas se asocian a un color. La primera columna muestra los resultados de segmentación mientras que en la segunda los resultados de postprocesamiento. De esta manera se observa que el postprocesamiento logra separar mejor las regiones fusionadas para la imagen de alta y media calidad. Sin embargo, para la imagen de baja calidad el resultado no es óptimo y por consiguiente se ve afectado cuando hay una excesiva concentración de células superpuestas, lo cual aumenta los errores al separarlas.



(a) (b)

Figura 4.6: Resultados de postprocesamiento para separar regiones en imágenes, en orden descendente por renglón: alta, media y baja calidad. a) I_{seg} con etiquetas. b) Transformación de *watershed*.

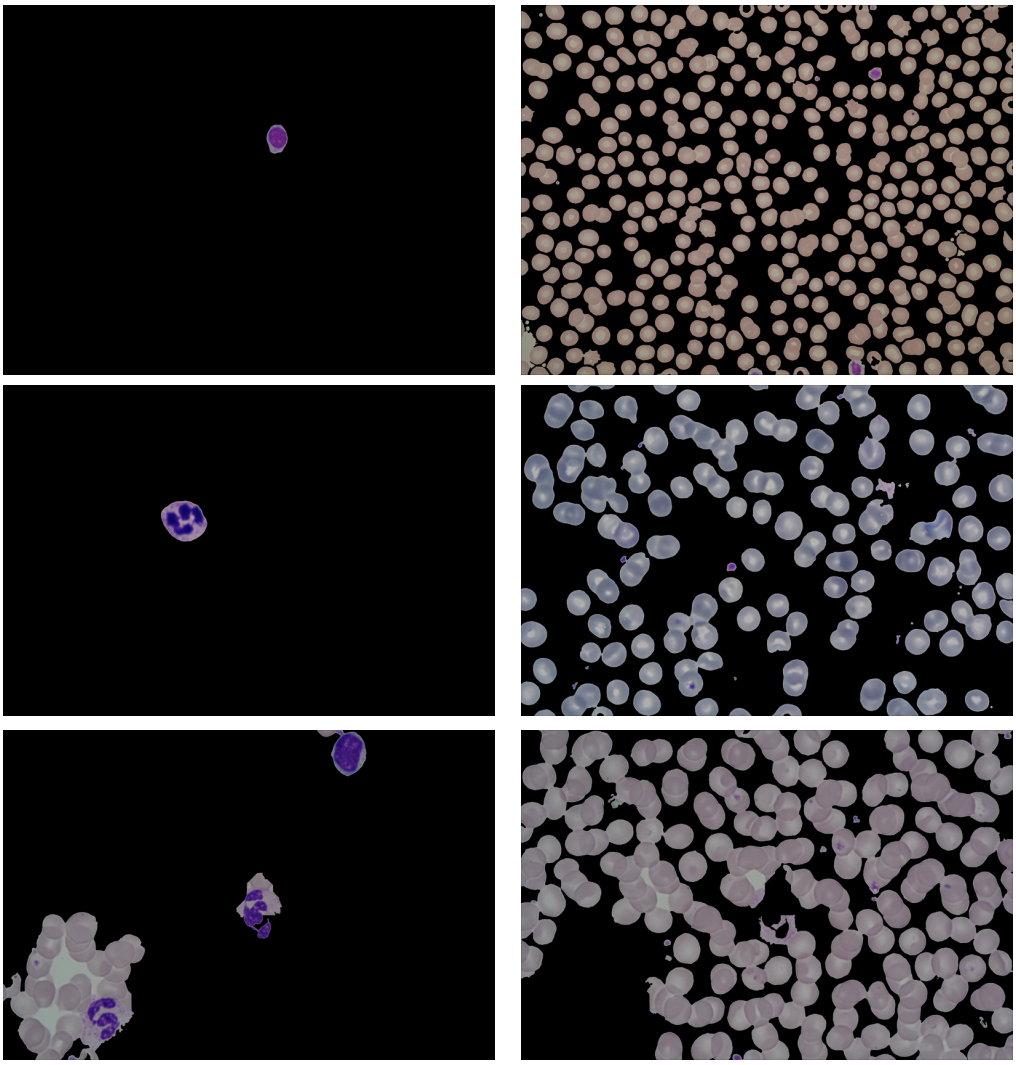
4.2.3. Identificación y conteo de leucocitos y eritrocitos

En esta subsección se presentan los resultados obtenidos de la identificación de leucocitos y eritrocitos, aplicados al conjunto de imágenes de frotis sanguíneo seleccionadas en la Sección 4.1. Dichos resultados se presentan considerando el resultado visual y aspectos comparativos con la identificación realizada por el especialista.

Los resultados mostrados para los tres tipos de imágenes son generados siguiendo el procedimiento descrito en la Subsección 3.2.1. Este método propuesto funciona bajo el supuesto de que en la imagen de frotis sanguíneo está presente al menos un leucocito. Además, el leucocito tiene que ser visible para el ojo humano y tener un citoplasma definido. Las imágenes mostradas en la Figura 4.1, aunque tienen diferente calidad en el frotis, cumplen con los criterios anteriores.

La Figura 4.7.a muestra resultados visuales de la identificación de leucocitos. Para la imagen de alta y media calidad, se identificaron todos los leucocitos marcados por el especialista. Para el caso en la imagen de baja calidad se identificaron los tres leucocitos marcados, pero la segmentación en uno de ellos fue deficiente, es decir, el leucocito fue subsegmentado a causa de la gran acumulación de eritrocitos a su alrededor. Este fenómeno ocurre generalmente en imágenes que no fueron adquiridas en la zona de lectura de frotis ideal. La Figura 4.7.b muestra los resultados visuales de la identificación de eritrocitos, esta imagen no debe incluir leucocitos, ya que previamente se extrajeron, resultando una imagen que contiene solo eritrocitos.

Los resultados visuales obtenidos muestran que esta propuesta identifica mejor los leucocitos en imágenes de calidad alta y media. En el caso de la imagen de calidad baja la identificación de leucocitos es deficiente debido al factor de células superpuestas cercanas a la célula, generando complejidad al tratar de recuperar el leucocito. Lo anterior puede generar una sobresegmentación o una subsegmentación en los leucocitos. Para el caso de los eritrocitos, éstos se identifican con el proceso de ir eliminado a leucocitos obtenidos sobre la imagen resultante en la Subsección 4.2.2.

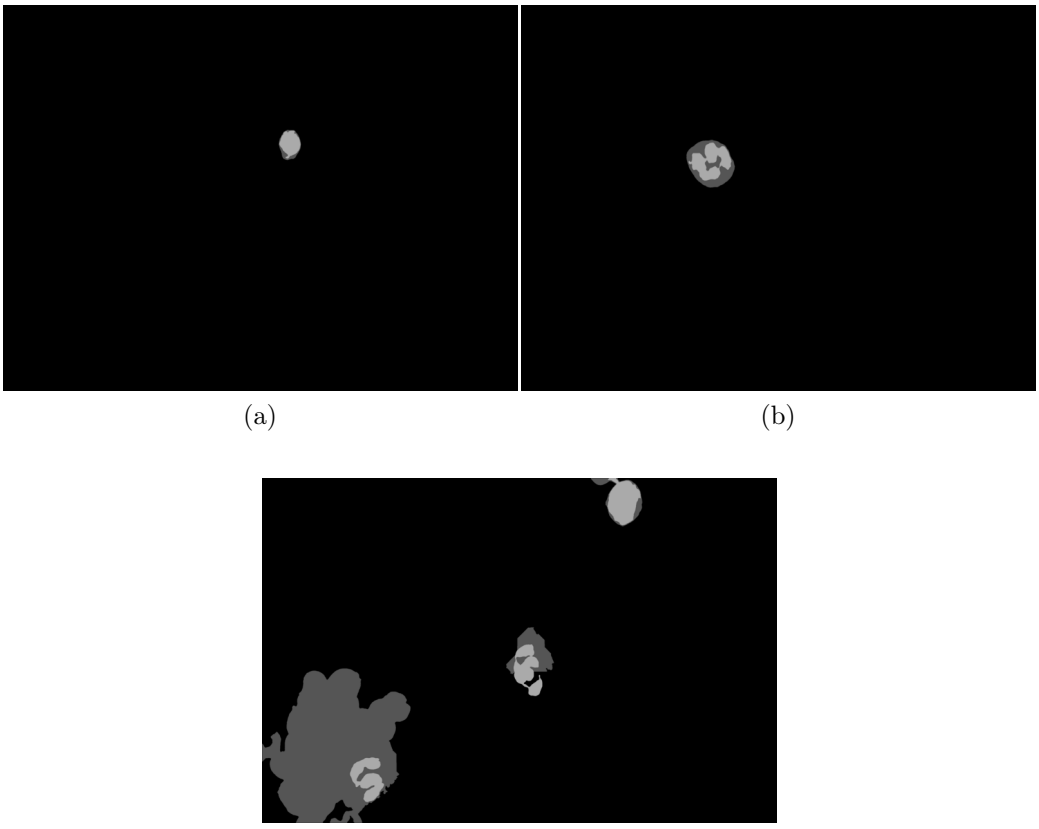


(a) (b)

Figura 4.7: Resultados de identificación de eritrocitos y leucocitos en imágenes, en orden descendente por renglón: alta, media y baja calidad. a) Leucocitos. b) Eritrocitos.

Como fue descrito en el módulo de "Identificación de leucocitos y eritrocitos", este proceso genera dos máscaras de segmentación, una contiene los eritrocitos y la otra los leucocitos. Entonces, se procede a generar la máscara de segmentación conjunta de leucocitos, mediante la unión de la máscara de

segmentación de núcleos y leucocitos, como se mostró en la Figura 3.6. La Figura 4.8 muestra la máscara de segmentación conjunta obtenida para las imágenes de calidad alta, media y baja.



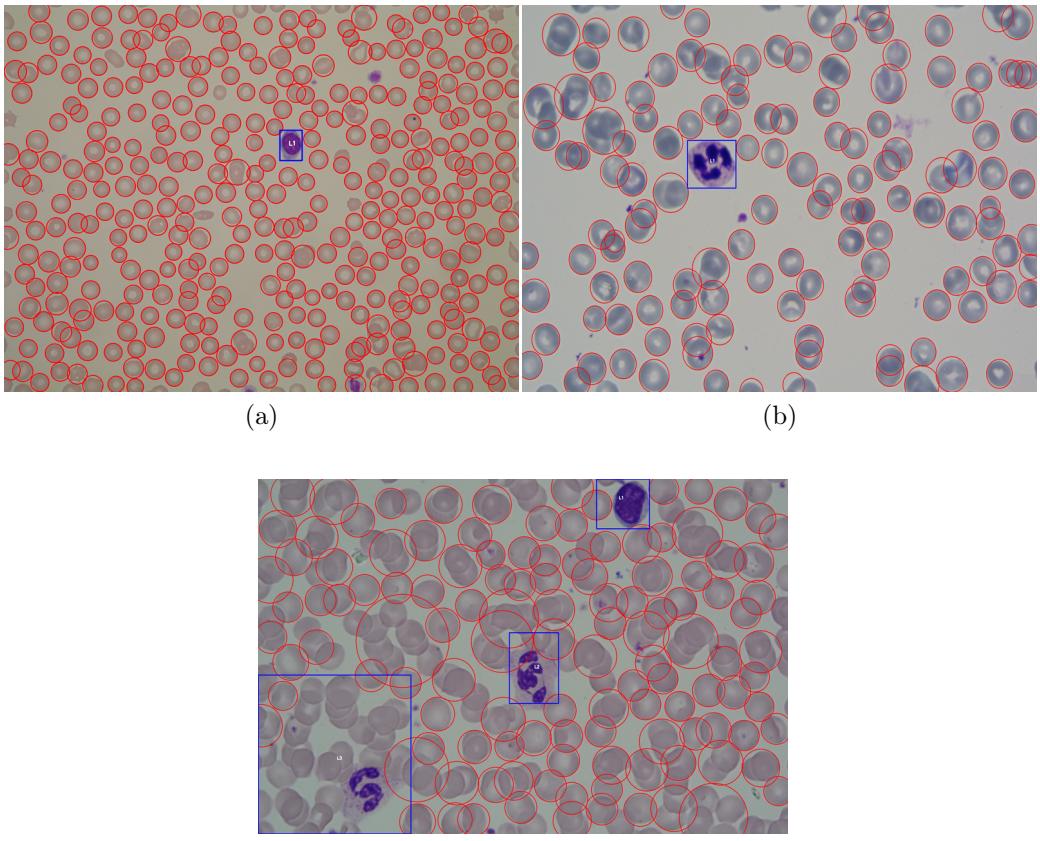
(a) (b)

(c)

Figura 4.8: Máscara de segmentación conjunta para la imagen de a) alta, b) media y c) baja calidad.

Una vez identificados los eritrocitos y leucocitos, se realiza el conteo automático de éstos. En primer lugar, se cuentan las regiones en la imagen de eritrocitos usando el método propuesto para el conteo automático que se describe en la Subsección 3.2.1. Para realizar el conteo automático de eritrocitos, estas regiones deben cumplir dos restricciones. La primera restricción propuesta en [Acharya and Kumar, 2018] establece que el factor de forma de la región debe estar en el rango $(0.5, 1)$. La segunda restricción establece que la región debe ser mayor en área a un percentil P_{15}, es decir, ser mayor al 15

porciento de las observaciones de área. En caso de que una región no cumpla con las dos restricciones es descartada para el conteo. A continuación, las regiones de la máscara de segmentación de leucocitos son etiquetadas, así el número de etiquetas corresponde con el conteo de leucocitos. La Figura 4.9 representa los resultados de conteo dibujando círculos para los eritrocitos y rectángulos para los leucocitos.



(a) (b)

(c)

Figura 4.9: Resultados de conteo de leucocitos y eritrocitos en imágenes de a) alta, b) media y c) baja calidad.

De las regiones candidatas a leucocitos mostradas en la Figura 4.7.a se generan subimágenes con el objetivo de ser utilizadas para generar un modelo para el reconocimiento de leucocitos. Entonces con ayuda de la máscara de segmentación conjunta se generan subimágenes de los leucocitos identificados, esto es, se obtienen las coordenadas del rectángulo más pequeño que contiene

al leucocito. Estas coordenadas sirven para recortar simultáneamente a los leucocitos en la imagen de frotis sanguíneo y su correspondiente en la máscara de segmentación conjunta. En la Figura 4.10 se pueden observar imágenes de los leucocitos identificados y la máscara de segmentación para la imagen de alta, media y baja calidad. Los leucocitos presentados en las columnas uno y dos corresponden a imágenes de frotis con calidad alta y media, mientras que las columnas tres, cuatro y cinco corresponden a uno de calidad baja. Aunque visualmente los núcleos de los leucocitos se logran segmentar correctamente para los tres tipos de calidad de frotis, en el caso del citoplasma, éste no se logra segmentar adecuadamente cuando la calidad del frotis es baja.

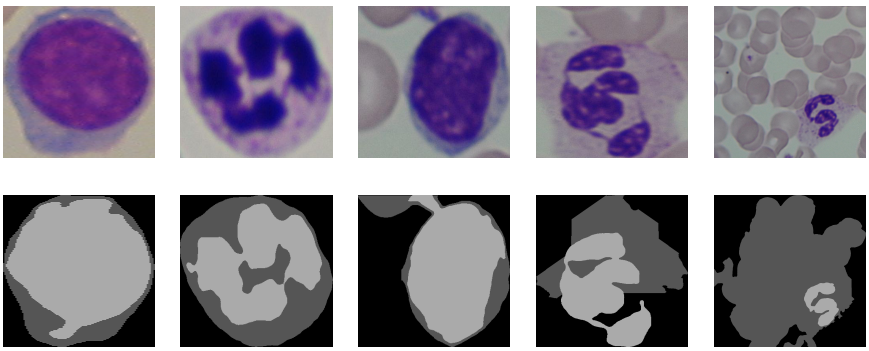


Figura 4.10: Imágenes de leucocitos aislados (primera fila) y máscaras de segmentación (segunda fila).

Núm. imágenes	**Conteo manual**	**Conteo automático**	
		Leucocitos	**Ruido**
149	219	209	19
Total	219	228	

Cuadro 4.5: Conteo de regiones candidatas a leucocitos en imágenes: comparativa de conteo manual vs conteo automático.

Más allá de los resultados visuales de la identificación y conteo en las imágenes de prueba, una manera cuantitativa de medir el conteo obtenido de manera automática es mediante una comparación con el conteo dado por el especialista. El Cuadro 4.5 muestra los resultados del conteo realizado por el especialista y el conteo automático para un conjunto de 149 imágenes de

frotis sanguíneo que contiene 219 leucocitos, el cual fue descrito en la Sección 4.1 y resultado de la unión de DS_1 y DS_2.

De los resultados de conteo obtenidos se observa que el conteo automático logra contar correctamente 209 candidatos a leucocitos, es decir, el proceso de identificación de leucocitos logra detectar de las 149 imágenes de frotis sanguíneo hasta un 95.43 % de los leucocitos.

Analizando los resultados, se observa que debido al tamaño y color de las megaplaquetas u otro elemento con las mismas características a un leucocito hacen que el proceso de identificación se equivoque y por consecuencia contabiliza dichos hallazgos como leucocitos. Cabe mencionar que estos hallazgos son casos atípicos en el frotis por lo que no se presentan con frecuencia. Otra observación es que algunos linfocitos no tienen el citoplasma definido y éste tiende a confundirse con el fondo de imagen, entonces se vuelve difícil segmentar, generando una sobregmentación. Por ejemplo, la Figura 4.11 ilustra una imagen de frotis sanguíneo con un neutrófilo que carece de una buena tinción en el citoplasma y sólo son visibles sus núcleos. Por el contrario, existen algunos casos donde hay células excesivamente acumuladas alrededor de los linfocitos generando una subsegmentación (ver Figura 4.6).

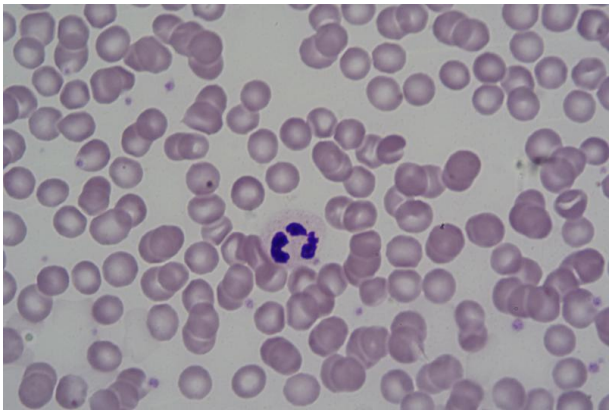


Figura 4.11: Neutrófilo con baja calidad de tinción en el citoplasma.

Respecto a la comparativa del conteo de eritrocitos, debido a la complejidad de dicha tarea, el especialista realizó este conteo sólo para 18 imágenes.

El Cuadro 4.6 muestra la comparación entre el conteo manual y el conteo usando el método propuesto.

Conjunto de imágenes	Imagen	Conteo manual	Conteo automático
DS_1	F_1.jpg	407	410
	F_15.jpg	373	366
	F_22.jpg	372	362
	F_24.jpg	287	323
	F_25.jpg	354	349
	F_26.jpg	334	329
	F_27.jpg	357	374
	F_28.jpg	383	395
	F_32.jpg	306	326
	F_33.jpg	374	356
	F_36.jpg	326	320
	F_37.jpg	312	310
DS_2	F_1.jpg	143	137
	F_42.jpg	151	152
	F_43.jpg	152	140
	F_52.jpg	113	125
	F_55.jpg	138	148
	F_76.jpg	120	122
	F_77.jpg	155	141
	Total	5295	5317

Cuadro 4.6: Conteo de eritrocitos: comparativa de conteo manual vs conteo automático.

4.3. Reconocimiento de leucocitos

En esta sección se describen las pruebas realizadas y resultados de clasificación de leucocitos considerando el enfoque de clasificación con los cinco principales leucocitos y por etapas (tipo y subtipo), usando los dos enfoques planteados en la Subsección 4.3.1. Es decir, los experimentos siguen la metodología descrita en la Figura 1.1 con los tres módulos presentados en la

Subsección 3.2.2. La primera fase tiene por objetivo extraer las características a las imágenes de leucocitos y así generar el conjunto de datos que servirá para obtener los modelos de clasificación de los leucocitos acorde con los enfoques planteados. La segunda fase se enfoca en obtener el mejor modelo de clasificación de leucocitos.

4.3.1. Enfoques de clasificación de leucocitos

El reconocimiento de leucocitos en este proyecto de tesis se ha realizado utilizando dos enfoques de clasificación: a) con cinco clases y b) tipos principales y subtipos de granulocitos. Esto es derivado de utilizar la taxonomía de los leucocitos mencionada en el Anexo A.2.

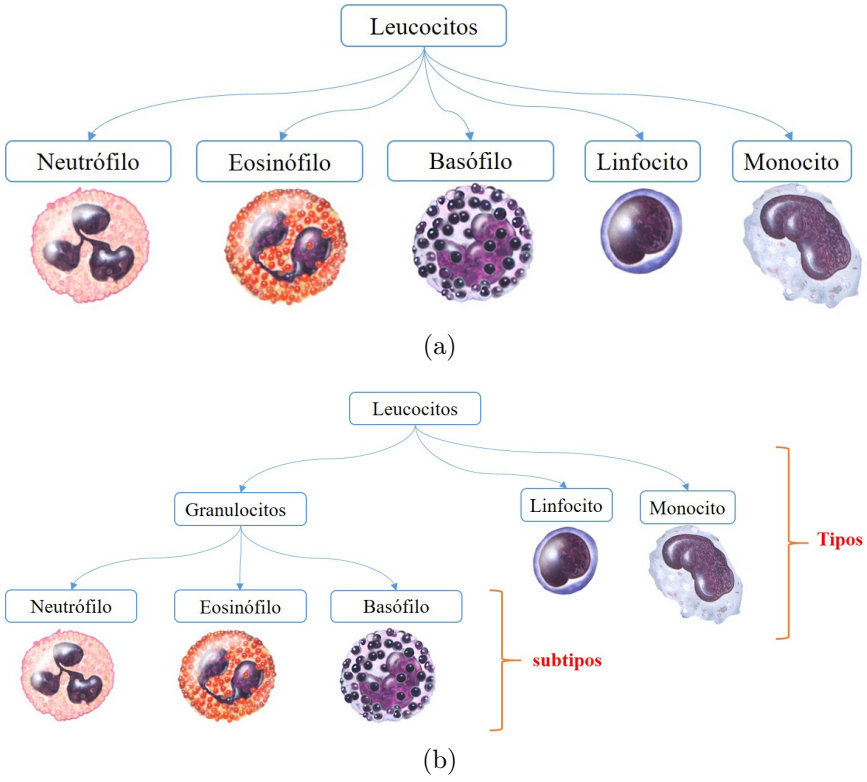


Figura 4.12: Enfoques de clasificación de leucocitos. Clasificación mediante a) cinco clases principales y b) por tipos y subtipos de leucocitos.

El primer enfoque con cinco clases, considera realizar la clasificación en basófilo, eosinófilo, linfocito, monocito y neutrófilo. El segundo enfoque realiza una clasificación en dos etapas, donde la primera de ellas se encarga de la clasificación por tipos principales de leucocitos (linfocitos, monocitos y granulocitos). Una vez que se ha realizado esta clasificación, la segunda etapa se lleva a cabo si la clase del leucocito corresponde a un granulocito. La clasificación en este caso es respecto de los subtipos de leucocitos (basófilo, eosinófilo o neutrófilo). La Figura 4.12 muestra los dos esquemas de clasificación de leucocitos.

4.3.2. Extracción de características

En esta parte se presenta la descripción del conjunto de datos a utilizar para realizar las pruebas de la metodología planteada en la Figura 1.1. Para esto se utilizan los leucocitos identificados y segmentados de la etapa anterior junto con el conjunto de imágenes $Union_A$ descrito en la Sección 4.1 generando el nuevo conjunto de imágenes DS_{527}. La distribución del conjunto de imágenes DS_{527} está descrita en el Cuadro 4.7.

	$Union_A$	Segmentación Automática	**Total**
Núm. imágenes de leucocitos	317	210	527

Cuadro 4.7: Distribución de ejemplos por conjuntos de imágenes.

Al conjunto de imágenes de leucocitos DS_{527}, se le aplica el proceso mostrado en la Figura 3.10. Para lograr la extracción de características se requiere aislar las regiones de interés del leucocito, en la Figura 3.11 se ilustran. A continuación se calculan las características para el núcleo, citoplasma y la célula para cada una de las imágenes de leucocitos, generando a CSV_{527}, un conjunto de datos de 527 ejemplos de leucocitos y 193 características. Dichas características corresponden a las descritas en el Cuadro 3.6.

El conjunto de datos generado anteriormente se utiliza para dividirlo en el conjunto de entrenamiento y prueba para cada uno de los enfoques de clasificación descritos en la Subsección 4.3.1. Para los enfoques de clasificación de cinco clases, tipo y subtipo, la distribución de los datos es mostrada en los Cuadros 4.8-4.10.

Célula	Etiqueta	Núm. ejemplos
Basófilo	C1	21
Eosinófilo	C2	36
Linfocito	C3	188
Monocito	C4	97
Neutrófilo	C5	185
	Total	527

Cuadro 4.8: Distribución de leucocitos, respecto al enfoque de clasificación con cinco clases.

Tipo	Etiqueta	Núm. ejemplos
Granulocito	C1	242
Linfocito	C3	188
Monocito	C4	97
	Total	527

Cuadro 4.9: Distribución de leucocitos, respecto al enfoque de clasificación en dos etapas: tipos.

subtipo	Etiqueta	Núm. ejemplos
Basófilo	C1	21
Eosinófilo	C2	36
Neutrófilo	C5	185
	Total	242

Cuadro 4.10: Distribución de leucocitos, respecto al enfoque de clasificación en dos etapas: subtipos.

4.3.3. Resultados de reconocimiento de leucocitos

En esta sección se describen los pasos para obtener el mejor modelo de clasificación de leucocitos. Primero evaluamos seis algoritmos de clasificación

para determinar los tres más adecuados para el problema de reconocimiento de leucocitos. En segundo lugar se realiza la selección de características a través de tres métodos y se evalúan estos subconjuntos de características con los tres algoritmos de clasificación obtenidos anteriormente. Estos algoritmos son optimizados mediante una búsqueda de parámetros para generar los mejores modelos con estas características.

Selección de los modelos de clasificación

Como primer paso en el proceso de obtener el mejor modelo de clasificación de leucocitos, se evalúan y comparan seis clasificadores. El objetivo de esta evaluación es obtener un panorama del rendimiento de los clasificadores y seleccionar a tres que tengan los mejores rendimientos, considerando tres métricas para evaluar los modelos. Para las pruebas se utiliza el conjunto de datos CSV_{527} obtenido en la Subsección 4.3.2. Este conjunto de datos es normalizado a modo que todos los valores de las características estén en el rango $[0, 1]$ y generar el conjunto normalizado $CSVN_{527}$. El conjunto de datos es separado de manera aleatoria y estratificada en cinco clases, con un 80 % para entrenamiento y 20 % para prueba.

A continuación, los ejemplos de entrenamiento y prueba para los enfoques de clasificación son los mismos, es decir, los ejemplos serán los mismos independientemente del enfoque de clasificación. La Figura 4.13 ilustra el proceso de creación del conjunto de datos para el entrenamiento y prueba para el segundo enfoque a partir de los datos del primer enfoque. La Figura 4.13.b ilustra que la unión de los ejemplos con clases C1, C2 y C5 (basófilo, eosinófilo y neutrófilo) genera la clase C1' (granulocitos), esto indica que los ejemplos son los mismos solo que su etiqueta cambia a C1', las etiquetas de las clases sobrantes son las mismas. Finalmente, para generar el conjunto de entrenamiento y prueba para los subtipos, se usan los ejemplos de la clasificación con 5 clases pero quitando a los ejemplos con clase C3 y C4 (ver Figura 4.13.c).

Para los experimentos, se utilizaron los siguientes clasificadores: máquinas de soporte vectorial (con núcleo lineal y con núcleo gaussiano), *k-nn*, bayesiano ingenuo (*Naive Bayes*), árbol de decisión (*Decision Tree*) y bosques aleatorios (*Random Forest*).

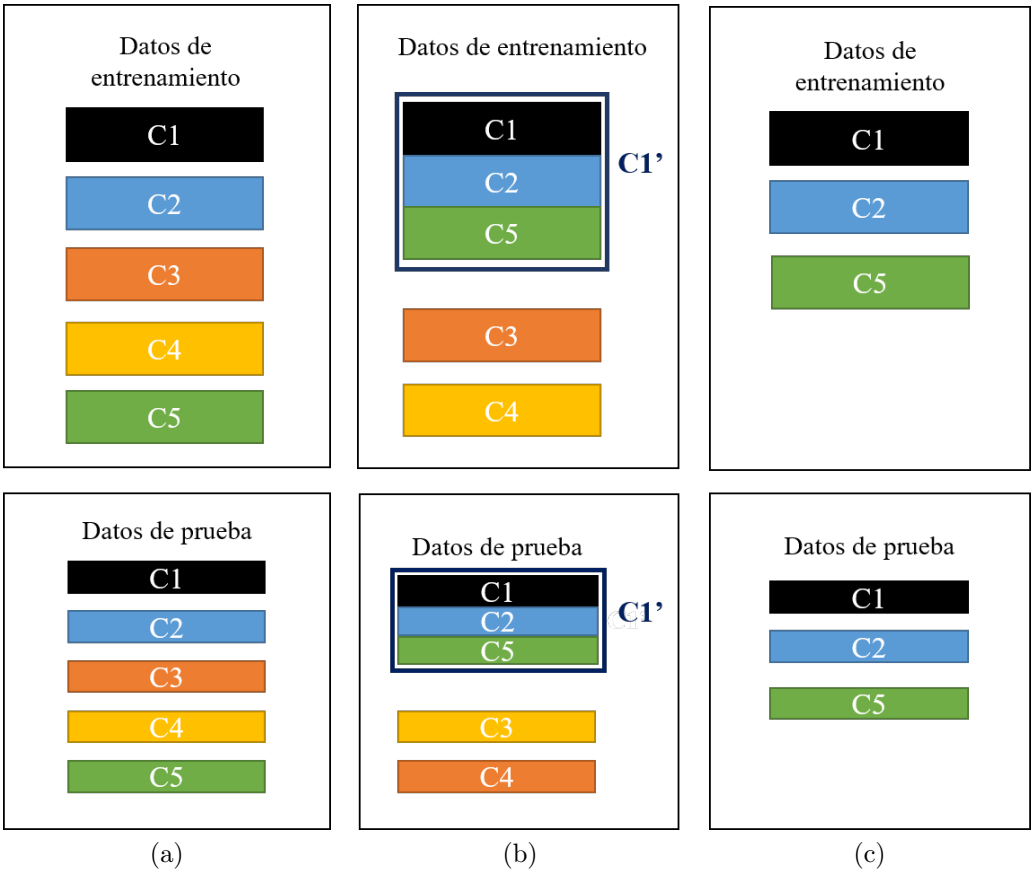


Figura 4.13: Generación de datos de entrenamiento y prueba: conjunto de datos para el enfoque con 5 clases, tipo (3 clases) y subtipo (3 clases) a) DS_{5C}, b) DS_{Tipo} y c) $DS_{subtipo}$. El símbolo C1' significa la agrupación de tres clases en el tipo granulocito.

Las pruebas realizadas en esta fase se describen a continuación. Primero, se entrenan seis clasificadores para los tres modos de clasificación de leucocitos y se realiza el ajuste (*tunnig*) de parámetros mediante una búsqueda en malla (*grid search*). Dicho de otra manera, se realiza una búsqueda de parámetros que maximicen la exactitud del modelo. En la Tabla 4.11 se muestra una descripción general de los parámetros considerados, el intervalo de búsqueda de los parámetros y sus incrementos para realizar una búsqueda de malla en cada clasificador.

Clasificador	Parámetros	Valores	Incrementos
SVM (Lineal)	C	$(2^{-5}, 2^{15})$	1 (exponente)
SVM (RBF)	C	$(2^{-5}, 2^{15}$	1 (exponente)
	γ	$(2^{-15}, 2^{3})$	1 (exponente)
k-nn	Métrica de distancias	Euclidiana, Cityblock, Minkowski	-
	Pesos	Uniforme, Distancia	-
	k vecinos	(1,21)	3
Decision Tree	Máxima profundidad	(5,50)	11
	max_features	$[None, sqrt, log2]$	-
	min_samples_leaf	$(2, 10)$	2
	Criterio	$[gini, entropy]$	-
Random Forest	Máxima profundidad	(5,50)	11
	max_features	$[None, sqrt, log2]$	-
	min_samples_leaf	$(2, 10)$	2
	Criterio	$[gini, entropy]$	-
	n_estimators	(500,2000)	100
Naive Bayes	-	-	-

Cuadro 4.11: Parámetros de los algoritmos de clasificación.

Para cada clasificador, se realiza un búsqueda de malla exhaustiva, donde cada combinación de parámetros son validados utilizando validación cruzada (*cross-validation)* con 10 particiones sobre el conjunto de entrenamiento. Al final nos quedamos con la mejor combinación de parámetros que nos devuelve esta búsqueda.

Una vez que se encuentran los parámetros óptimos que maximizan la exactitud del modelo, se calculan las medidas de evaluación sobre el conjunto de prueba para determinar qué tan bien generaliza el modelo [Tharwat, 2018]. Estos son el coeficiente de correlación de Matthews (MCC, por sus siglas en inglés), tasa de error balanceada (BER, por sus siglas en inglés), la exactitud balanceada (BAS, por sus siglas en inglés) y la exactitud (*accuracy*).

En los Cuadros 4.12-4.14 se muestran los resultados obtenidos de cada enfoque de clasificación de leucocitos y el proceso utilizado para la generación de los modelos es mostrado en la Figura 3.14

Clasificador	**MCC**	**BER**	**BAS**	**Exactitud**
SVM (Lineal)	0.8806	0.1301	0.8699	0.9150
SVM (RBF)	0.8806	0.1301	0.8699	0.9150
Random Forest	0.8275	0.2731	0.7269	0.8773
k-nn	0.8116	0.2758	0.7242	0.8679
Decision Tree	0.7379	0.3598	0.6402	0.8113
Naive Bayes	0.7271	0.3343	0.6657	0.8018

Cuadro 4.12: Resultados de medidas de evaluación de los modelos para el reconocimiento de leucocitos para el caso cuando la distribución de clases es en basófilo, eosinófilo, linfocito, monocito y neutrófilo.

Clasificador	**MCC**	**BER**	**BAS**	**Exactitud**
SVM (RBF)	0.9256	0.07544	0.9246	0.9528
Random Forest	0.9104	0.08238	0.9176	0.9433
SVM (Lineal)	0.8957	0.1009	0.8991	0.9339
Decision Tree	0.8662	0.1184	0.8816	0.9150
k-nn	0.8514	0.1102	0.8898	0.9056
Naive Bayes	0.8107	0.1286	0.8714	0.8773

Cuadro 4.13: Resultados de medidas de evaluación de los modelos para el reconocimiento de leucocitos para el caso cuando la distribución de clases es por tipo, es decir, linfocito, monocito y granulocito.

Clasificador	**MCC**	**BER**	**BAS**	**Exactitud**
SVM (Lineal)	0.724	0.1966	0.8034	0.8958
Decision Tree	0.7156	0.2352	0.7648	0.8958
SVM (RBF)	0.7017	0.3452	0.6548	0.8958
k-nn	0.6405	0.3571	0.6429	0.875
Random Forest	0.6334	0.3929	0.6071	0.875
Naive Bayes	0.5024	0.4199	0.5801	0.8125

Cuadro 4.14: Resultados de medidas de evaluación de los modelos para el reconocimiento de leucocitos para el caso cuando la distribución de clases es por subtipo, es decir, basófilo, eosinófilo, y neutrófilo.

En el caso del índice MCC y considerando los resultados obtenidos por las SVM (tanto lineal como RBF) y *Random Forest*, para una clasificación con

cinco clases, el valor en estos clasificadores se encuentra por arriba del 0.8206, alcanzando su máximo valor con las SVM. Asimismo, para una clasificación por tipos, el valor de éstos se encuentran por arriba de 0.89, alcanzando su máximo valor con la *SVM (RBF)*. En contraste, la clasificación por subtipo, alcanzó su máximo valor en 0.724 con una *SVM (Lineal)*.

Respecto al BER, en el caso de las SVM (tanto lineal como RBF) en los Cuadros 4.12 y 4.13 se encuentran por debajo de 0.13 y para el Cuadro 4.14 se encuentra por debajo de 0.34. Es decir, el error por clases para la clasificación con 5 clases y tipos es pequeño comparado con el de subtipos.

Los valores del BAS son importantes debido a que este índice es para conjuntos de datos desbalanceados. Al analizar los resultados del Cuadro 4.12, para las SVM (tanto lineal como RBF) el BAS es de 0.8699, por lo tanto se considera a las SVMs como un buen algoritmo de aprendizaje supervisado para resolver el problema de clasificación con un enfoque con 5 clases. Para el caso de clasificación por tipos, se observa que la *SVM (RBF)* y *Random Forest* el valor del BAS en estos se encuentra por arriba de 0.91. En el Cuadro 4.14 se observa que el BAS alcanza su máximo valor 0.8034 con *SVM (Lineal)*.

Combinado la información proporcionada por el MCC, BER y BAS se concluye que las SVM (tanto lineal como RBF) y *Random Forest* son ideales para el enfoque de clasificación con 5 clases y por tipos. En contraste, en el Cuadro 4.14 se observa que solo la *SVM (Lineal)* obtuvo buenos resultados respecto al BAS y BER comparado con el segundo mejor de esta tabla. Por lo anterior, en la clasificación por subtipos la *SVM (Lineal)* es mejor que los otros clasificadores.

Con estos resultados podemos concluir que *Random Forest* y las SVM (tanto lineal como RBF) son mejores para el reconocimiento de leucocitos que los otros clasificadores utilizados en estas pruebas. Cabe destacar que las pruebas utilizaron las 193 características calculadas originalmente.

Selección de características y ajuste de parámetros

En este apartado se presentan los resultados de realizar la selección de características. Esto es con el objetivo de obtener una representación de los leucocitos más reducido, es decir, generar un subconjunto de características que ofrezca un mejor rendimiento al original. Para realizarlo se aplica el procedimiento mostrado en la Figura 3.12.

Primero se emplean dos métodos para optimizar los modelos de clasificación: a) se reduce el número de variables usando eliminación recursiva de características con validación cruzada (RFECV, por sus siglas en inglés), selección de características basada en un umbral (TFS) y la selección de características univariadas (UFS) y b) se generan nuevos modelos de clasificación con esas características, éstos modelos serán optimizados mediante el ajuste parámetros. Para mayor información sobre los métodos véase Subsección 2.3.3.

La selección de características univariadas selecciona las mejores características basadas en una prueba estadística univariada. Al elegir como función de puntuación la información mutua (MI), la cual mide la dependencia entre las variables y las etiquetas para seleccionar las k mejores características [Wei and Stocker, 2016]. El proceso de selección de características para este método es mostrado en la Figura 3.12b. Aquí, el rango de valores para k es de $(10, 150)$, lo cual genera un total de 139 conjuntos de datos con k características. Cada conjunto de datos es evaluado, es decir, se generan 139 modelos (ver Figura 3.13) lo cual es realizado siguiendo el proceso mostrado en la Figura 3.14. Al término de la evaluación de los modelos generados, se selecciona el mejor considerando el mayor valor en MCC, menor para el BER y mayor BAS, entonces se almacenan cuántas y cuáles características fueron usadas en el proceso de entrenamiento.

El segundo método de selección de características se basa en RFE, el cual selecciona características considerando recursivamente conjuntos de características cada vez más pequeños. Primero, se entrena un clasificador sobre el conjunto de entrenamiento para obtener la importancia de las características. Luego, las características cuyos valores de importancia son los más pequeños se eliminan. Este procedimiento se repite recursivamente. RFECV realiza RFE con validación cruzada para encontrar un conjunto óptimo de características. Como RFECV requiere un clasificador para obtener la importancia de las características, se utiliza bosques aleatorios con 10 árboles.

El método de selección de características basada en un umbral requiere que la importancia de las características sean calculadas. Por lo anterior, se entrenó el modelo basado en árboles extremadamente aleatorios (*Extremely Randomized Trees*) utilizando 700 árboles. Entonces, las características se seleccionan con una importancia mayor o igual al valor umbral $t = 0.007$. De este modo, aquellas características que sean inferiores al valor umbral se

eliminan. El modelo, sus parámetros y el umbral a utilizar fueron obtenidos de manera experimental.

Los Cuadros 4.15-4.17 muestran los resultados obtenidos al aplicar estos métodos de selección de características para los tres enfoques de clasificación de leucocitos propuestos.

En el Cuadro 4.15 se observa que de los 3 métodos de selección de características, el que alcanza un valor máximo en MCC, BAS y mínimo BER es el método TFS, el cual selecciona 49 de las 193 características. Al comparar el rendimiento de este modelo con el que utiliza 193 características (ver Cuadro 4.12), se obtiene que con 49 características y usando *SVM (Lineal)* el rendimiento es superior. Por lo tanto, el mejor modelo de clasificación para el enfoque de 5 clases es la *SVM (Lineal)* utilizando 49 características. El parámetro de este modelo es $C = 128$.

Método	**Núm. Características**	**Clasificador**	**MCC**	**BER**	**BAS**	**Exactitud**
UFS	**60**	*SVM (Lineal)*	0.9071	0.07465	0.9254	0.9339
		SVM (RBF)	0.8933	0.1432	0.8568	0.9245
		Random Forest	0.8538	0.2364	0.7636	0.8962
RFECV	**98**	*SVM (Lineal)*	0.8946	0.1125	0.8875	0.9245
		SVM (RBF)	0.8693	0.1492	0.8508	0.9056
		Random Forest	0.8545	0.281	0.719	0.8962
TFS	**49**	*SVM (Lineal)*	0.9342	0.05479	0.9452	0.9528
		SVM (RBF)	0.9076	0.1471	0.8529	0.9339
		Random Forest	0.8538	0.2364	0.7636	0.8962

Cuadro 4.15: Resultados de selección de características y medidas de evaluación de los modelos para el reconocimiento de leucocitos para cinco clases.

En el Cuadro 4.16 se observa que en el caso del índice MCC y considerando solo los resultados de la SVM (tanto lineal como RBF), el valor en los tres métodos se encuentra por arriba del 0.9251, alcanzado su valor máximo en 0.9557 con *SVM (Lineal)* utilizando solo 49 características. La otra medida de evaluación es el índice BER, el valor mínimo alcanzado es 0.047 con el método TFS que selecciona 47 de las 193 características. Para el BAS, el valor máximo obtenido es 0.95 con los métodos UFS y TFS. Con estos resultados podemos concluir que el mejor modelo para reconocer leucocitos utilizando el enfoque

de clasificación por tipos es la *SVM (Lineal)* utilizando 49 características. El parámetro de este modelo es $C = 32$.

Método	**Núm. Caracte-ríticas**	**Clasificador**	**MCC**	**BER**	**BAS**	**Exactitud**
UFS	**49**	*SVM (Lineal)*	0.9557	0.05	0.95	0.9716
		SVM (RBF)	0.9251	0.05599	0.944	0.9528
		Random Forest	0.8953	0.09905	0.901	0.9339
RFECV	**7**	*SVM (Lineal)*	0.9412	0.06667	0.9333	0.9622
		SVM (RBF)	0.9256	0.07361	0.9264	0.9528
		Random Forest	0.8951	0.09115	0.9088	0.9339
TFS	**47**	*SVM (Lineal)*	0.9402	0.04722	0.9528	0.9622
		Random Forest	0.9261	0.07544	0.9246	0.9528
		SVM (RBF)	0.9254	0.07361	0.9264	0.9528

Cuadro 4.16: Resultados de selección de características y medidas de evaluación de los modelos para el reconocimiento de leucocitos para tres clases: tipo.

Método	**Núm. Caracte-rísticas**	**Clasificador**	**MCC**	**BER**	**BAS**	**Exactitud**
UFS	**109**	*SVM (Lineal)*	0.8887	0.09524	0.9048	0.9583
		SVM (RBF)	0.7679	0.2619	0.7381	0.9166
		Random Forest	0.6334	0.3929	0.6071	0.875
RFECV	**150**	*SVM (Lineal)*	0.8887	0.09524	0.9048	0.9583
		SVM (RBF)	0.7017	0.3452	0.6548	0.8958
		Random Forest	0.6334	0.3929	0.6071	0.875
TFS	**44**	*SVM (Lineal)*	0.7722	0.1876	0.8124	0.9166
		SVM (RBF)	0.7722	0.1876	0.8124	0.9166
		Random Forest	0.7043	0.3452	0.6548	0.8958

Cuadro 4.17: Resultados de selección de características y medidas de evaluación de los modelos para el reconocimiento de leucocitos para tres clases: subtipo.

En el Cuadro 4.17 se observa que existen dos métodos de selección de características con la *SVM (Lineal)* que generan los mismos resultados de evaluación. En el caso del MCC el valor máximo registrado es 0.8887 y para el BAS, el valor máximo es 0.9048. En contraste, los bosques aleatorios para cualquier método de selección de características su valor máximo para BAS es 0.6548, es decir, tiene un mal rendimiento comparado con la *SVM (Lineal)*. Para el BER, en el caso de las *SVM (Lineal)* con más alto MCC y BAS el valor está por debajo de 0.095, es decir, el error por clases es muy pequeño. En consecuencia los dos métodos con el mejor rendimiento son ideales para hacer la clasificación por subtipos. Sin embargo, considerando el número mínimo de características elegimos el método UFS que devuelve el mínimo de características, esto es, sólo 109 de las 193 características originales.

Para el enfoque de clasificación por subtipos, se observa que se necesita de una mayor cantidad de características comparada con los enfoques por tipos o por cinco clases principales. Esto se deriva a causa de que los basófilos, eosinófilos y neutrófilos comparten muchos atributos morfológicos y se vuelve complejo diferenciarse entre ellos. Por consecuencia, se requieren de más descriptores para poder obtener un buen rendimiento de generalización.

En resumen, los mejores modelos de clasificación y métodos de selección de características son mostrados en el Cuadro 4.18. Cabe mencionar que los mejores resultados se obtuvieron usando *SVM (Lineal)* para cada enfoque de clasificación. Los métodos que mejores características seleccionan son el UTF y TFS. Para más detalles sobre las características seleccionadas para cada enfoque de clasificación de leucocitos véase el Anexo B.

Enfoque	**Método**	**Núm. Caracte-rísticas**	**MCC**	**BER**	**BAS**	**Exactitud**
5C	TFS	**49**	**0.9342**	**0.05479**	**0.9452**	0.9528
Tipo	UFS	**49**	**0.9557**	**0.05**	**0.95**	0.9716
	TFS	**47**	0.9402	0.04722	0.9528	0.9622
Subtipo	UFS	**109**	**0.8887**	**0.09524**	**0.9048**	0.9583
	RFECV	**150**	0.8887	0.09524	0.9048	0.9583

Cuadro 4.18: Resumen de los mejores modelos para los tres tipos de enfoque de clasificación.

Con el fin de analizar la contribución de cada clase en la clasificación general (ver Cuadro 4.18), se muestran las matrices de confusión de los tres mejores modelos de clasificación de leucocitos (ver Cuadros 4.19-4.21). Para el enfoque de clasificación usando los cinco principales leucocitos se observa que el clasificador *SVM (Lineal)* es bueno al reconocer perfectamente los basófilos y linfocitos, seguido de eosinófilos y monocitos, los cuales sólo tienen un ejemplo clasificado incorrectamente. Debido a que los neutrófilos y eosinófilos pertenecen al mismo tipo de leucocitos (granulocitos) éstos suelen ser ligeramente similares por lo que tienden a confundirse (ver Cuadro 4.19).

	Basófilo	**Eosinófilo**	**Linfocito**	**Monocito**	**Neutrófilo**
Basófilo	**4**	0	0	0	0
Eosinófilo	0	**6**	0	1	0
Linfocito	0	0	**38**	0	0
Monocito	0	0	0	**19**	1
Neutrófilo	0	2	0	1	**34**

Cuadro 4.19: Matriz de confusión de la SVM lineal con el enfoque de clasificación 5C y utilizando el método de selección de características TFS.

Después de analizar los Cuadros 4.20 y 4.21, se observa que al realizar la clasificación en tipos, se logran clasificar correctamente los granulocitos y linfocitos. Sin embargo, para los monocitos asigna 3 ejemplos a otras clases. Por otro lado, se observa que el enfoque de clasificación con 5 clases, tiene un buen rendimiento en general excepto para clasificar neutrófilos lo cual se mejora utilizando el enfoque por subtipos, ya que ahí se logran reconocer correctamente todos los neutrófilos. Además, los leucocitos que fueron más fáciles de reconocer fueron los basófilos y linfocitos sin importar el enfoque de clasificación.

	Granulocito	**Linfocito**	**Monocito**
Granulocito	48	0	0
Linfocito	0	**38**	0
Monocito	2	1	**17**

Cuadro 4.20: Matriz de confusión de la SVM lineal con el enfoque de clasificación por tipos y utilizando el método de selección de características UFS.

	Basófilo	**Eosinófilo**	**Neutrófilo**
Basófilo	4	0	0
Eosinófilo	0	**5**	2
Neutrófilo	0	0	**37**

Cuadro 4.21: Matriz de confusión de la SVM lineal con el enfoque de clasificación por subtipos y utilizando el método de selección de características UFS.

En resumen, se puede utilizar cualquiera de las dos propuestas de clasificación planteadas debido a ambas tienen buen rendimiento al reconocer los leucocitos. Derivado de lo anterior, por simplicidad el modelo seleccionado que formará parte de la aplicación web para reconocer leucocitos es el enfoque de clasificación con 5 clases que solo utiliza 49 características y *SVM (Lineal)* como clasificador.

Capítulo 5

Conclusiones y trabajo a futuro

En este proyecto de tesis se desarrollaron varios procedimientos para identificar y contar eritrocitos, y regiones candidatas a leucocitos a partir de imágenes microscópicas de frotis sanguíneo. Estos procedimientos abordan diversas etapas, de las cuales la más importante es la segmentación de las células sanguíneas debido a que estos resultados son cruciales para generar buenos resultados de identificación y conteo de leucocitos y eritrocitos. En esta fase se concluye que los resultados de identificación de células sanguíneas dependen directamente de la calidad del frotis sanguíneo. Es decir, a mayor concentración de células (aglutinamiento) en el frotis más complicado es la segmentación las células sanguíneas.

Para evaluar el conteo de eritrocitos y regiones candidatas a leucocitos se realizó una comparativa de estos contra el conteo generado por el especialista. Los experimentos mostraron que al identificar los núcleos leucocitarios e identificar las células sanguíneas se logran recuperar correctamente 209 regiones candidatas a leucocitos de un total de 219 registrados por el especialista. Para el caso de conteo de eritrocitos, los experimentos mostraron que el método de conteo al utilizar restricciones de forma y área, contabilizan 5317 eritrocitos de los 5295 registrados por el especialista.

Al realizar una evaluación cuantitativa de los modelos de clasificación y sus características seleccionadas, se obtuvo que el mejor algoritmo de aprendizaje supervisado para reconocer a los leucocitos es *SVM (Lineal)* independientemente del enfoque de clasificación. Los resultados obtenidos mostraron que el enfoque de clasificación con cinco clases, alcanzó el valor máximo en

exactitud, MCC, BAS y mínimo BER cuando se seleccionan 49 de las 193 características originales a través del método de selección de características por umbral. Por lo anterior, el modelo entrenado, utilizando 49 características, con una *SVM (Lineal)* es integrada a la aplicación web para reconocer leucocitos, en conjunción con los procedimientos para identificar y contar eritrocitos y leucocitos.

Debido a que las mediciones de MCC y BAS mostradas por los clasificadores aún podría mejorarse, se deja para un trabajo futuro utilizar otro tipo de características para describir mejor a los leucocitos y aplicar también otro tipo de métodos de selección de características. Además, otro proceso que se puede mejorar es la identificación de leucocitos, por lo que se podrían explorar métodos de aprendizaje autosupervisado para una rápida y robusta segmentación de leucocitos. Finalmente, se podría explorar la posibilidad de identificar y reconocer células anormales en el frotis, las cuales son precursoras de algún tipo de cáncer en la sangre.

Bibliografía

[Acharya and Kumar, 2018] Acharya, V. and Kumar, P. (2018). Identification and red blood cell automated counting from blood smear images using computer-aided system. *Medical & biological engineering & computing*, 56(3):483–489.

[AL-Dulaimi et al., 2018] AL-Dulaimi, K., Banks, J., Chandran, V., Tomeo-Reyes, I., and Nguyen Thanh, K. (2018). Classification of white blood cell types from microscope images: Techniques and challenges. In *Microscopy Science: Last Approaches on Educational Programs and Applied Research*, volume 8. Formatex Research Center.

[Aushev et al., 2018] Aushev, A., Ripoll, V. R., Vellido, A., Aletti, F., Pinto, B. B., Herpain, A., Post, E. H., Medina, E. R., Ferrer, R., Baselli, G., et al. (2018). Feature selection for the accurate prediction of septic and cardiogenic shock icu mortality in the acute phase. *PloS one*, 13(11):18.

[Ávarez et al., 2006] Ávarez, D. A., Guevara, M. L., and Holguín, G. A. (2006). Preprocesamiento de imágenes aplicadas a mamografías digitales. *Scientia et technica*, 12(31):1–6.

[Bishop, 2006] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

[Breiman, 2001] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

[Breiman et al., 1984] Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and Regression Trees*. Wadsworth.

[Burger and Burge, 2009] Burger, W. and Burge, M. J. (2009). *Principles of digital image processing*. Springer.

[Burges, 1998] Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, 2(2):121–167.

[Center, 2019] Center, M. R. (2019). Isfahan university of medical sciences. Recuperado mayo de 2019, de: http://misp.mui.ac.ir/fa/download.

[Chen et al., 1995] Chen, Y. Q., Nixon, M. S., and Thomas, D. W. (1995). Statistical geometric features for texture classification. *Pattern recognition*, 28(4):537–552.

[Cuevas et al., 2010] Cuevas, E., Osuna-Enciso, V., Oliva, D., Wario, F., and Zaldivar, D. (2010). Segmentación y detección de glóbulos blancos en imágenes usando sistemas inmunes artificiales. *Revista mexicana de ingeniería biomédica*, 31(2):119–134.

[Darst et al., 2018] Darst, B. F., Malecki, K. C., and Engelman, C. D. (2018). Using recursive feature elimination in random forest to account for correlated variables in high dimensional data. *BMC Genetics*, 19(1):65.

[Diem et al., 2004] Diem, H., Haferlach, T., and Theml, H. (2004). *Color Atlas of Hematology: Practical Microscopic and Clinical Diagnosis (Second Revised Edition)*. Thieme Publishing Group.

[Dougherty, 2013] Dougherty, G. (2013). *Pattern recognition and classification: an introduction*. Springer.

[Duda et al., 2000] Duda, R. O., Hart, P. E., and Stork, D. G. (2000). *Pattern Classification (Second Edition)*. Wiley-Interscience.

[Geurts et al., 2006] Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Machine learning*, 63(1):3–42.

[Gonzalez and Woods, 1996] Gonzalez, R. C. and Woods, R. E. (1996). *Tratamiento Digital de Imágenes)*. Addison Wesley Iberoamericana.

[Gonzalez and Woods, 2008] Gonzalez, R. C. and Woods, R. E. (2008). *Digital Image Processing (Third Edition)*. Pearson Prentice Hall.

[Gregorutti et al., 2017] Gregorutti, B., Michel, B., and Saint-Pierre, P. (2017). Correlation and variable importance in random forests. *Statistics and Computing*, 27(3):659–678.

[Guyon and Elisseeff, 2003] Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*, 3(3):1157–1182.

[Guyon et al., 2002] Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1-3):389–422.

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

[Hu, 1962] Hu, M.-K. (1962). Visual pattern recognition by moment invariants. *IRE Transactions on Information Theory*, 8(2):179–187.

[Huang and Hung, 2012] Huang, D.-C. and Hung, K.-D. (2012). Leukocyte nucleus segmentation and recognition in color blood-smear images. In *Proceedings of IEEE International Instrumentation and Measurement Technology Conference Proceedings*, pages 171–176.

[Isaza et al., 2018] Isaza, C., Anaya, K., De Paz, J. Z., Vasco-Leal, J. F., Hernandez-Rios, I., and Mosquera-Artamonov, J. D. (2018). Image analysis and data mining techniques for classification of morphological and color features for seeds of the wild castor oil plant (ricinus communis l.). *Multimedia Tools and Applications*, 77(2):2593–2610.

[Jaime Pérez and Gómez Almaguer, 2005] Jaime Pérez, J. C. and Gómez Almaguer, D. (2005). *Hematología: la sangre y sus enfermedades (Segunda Edición)*. McGraw Hill.

[Jiménez Díaz, 2007] Jiménez Díaz, L. (2007). *Clasificación de leucocitos mediante redes bayesianas*. Tesis de Licenciatura, Universidad Tecnológica de la Mixteca, México.

[Kim and Kim, 2000] Kim, W.-Y. and Kim, Y.-S. (2000). A region-based shape descriptor using zernike moments. *Signal processing: Image communication*, 16(1-2):95–102.

[Labati et al., 2011] Labati, R. D., Piuri, V., and Scotti, F. (2011). All-idb: The acute lymphoblastic leukemia image database for image processing.

In *2011 18th IEEE International Conference on Image Processing*, pages 2045–2048.

[Li et al., 2018] Li, Y., Cornelis, B., Dusa, A., Vanmeerbeeck, G., Vercruysse, D., Sohn, E., Blaszkiewicz, K., Prodanov, D., Schelkens, P., and Lagae, L. (2018). Accurate label-free 3-part leukocyte recognition with single cell lens-free imaging flow cytometry. *Computers in biology and medicine*, 96:147–156.

[Linder and Zahniser, 2012] Linder, J. and Zahniser, D. (2012). Digital imaging in hematology. *Medical Laboratory Observer*, 44(5):14–16.

[Longo, 2012] Longo, D. L. (2012). *Harrison: principios de medicina interna (18a.* McGraw Hill Mexico.

[López-Santiago, 2016] López-Santiago, N. (2016). Blood cytometry. *Acta pediátrica de México*, 37:246–249.

[Martínez Castro et al., 2014] Martínez Castro, J., Reyes Cadena, S., and Felipe Riverón, E. (2014). Leukocytes detection, classification and counting in smears of peripheral blood. *Revista Mexicana de Ingeniería Biomédica*, 35(1):41–51.

[MedlinePlus, 2018] MedlinePlus (2018). Problemas plaquetarios. Recuperado diciembre de 2018, de: https://medlineplus.gov/ spanish/plateletdisorders.html.

[Miao and Xiao, 2018] Miao, H. and Xiao, C. (2018). Simultaneous segmentation of leukocyte and erythrocyte in microscopic images using a marker-controlled watershed algorithm. *Computational and mathematical methods in medicine*, 2018:1–9.

[Moraleda Jiménez, 2017] Moraleda Jiménez, J. M. (2017). *Pregrado de Hematología (Cuarta Edición)*. Luzán5.

[Nixon and Aguado, 2012] Nixon, M. and Aguado, A. S. (2012). *Feature extraction and image processing for computer vision (Third Edition)*. Academic Press.

[Nodejs, 2019] Nodejs (2019). nodejs.org. Recuperado Junio de 2019, de: https://nodejs.org/es/download/.

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

[Pang et al., 2015] Pang, G., Zhuang, Y., and Zhou, P. (2015). Automatic leukocytes classification by distance transform, moment invariant, morphological features, gray level co-occurrence matrices and svm. In *First International Conference on Information Sciences, Machinery, Materials and Energy*, pages 1090–1095.

[Poynton, 1997] Poynton, C. (1997). Frequently asked questions about color. Recuperado octubre de 2018, de: http://poynton.ca/PDFs/ColorFAQ.pdf.

[Putzu and Di Ruberto, 2013] Putzu, L. and Di Ruberto, C. (2013). White blood cells identification and counting from microscopic blood image. In *Proceedings of the World Academy of Science, Engineering and Technology*, pages 15–22.

[Quinlan, 1986] Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, 1(1):81–106.

[Rezatofighi and Soltanian-Zadeh, 2011] Rezatofighi, S. H. and Soltanian-Zadeh, H. (2011). Automatic recognition of five types of white blood cells in peripheral blood. *Computerized Medical Imaging and Graphics*, 35(4):333–343.

[Rodak and Carr, 2015] Rodak, B. F. and Carr, J. H. (2015). *Clinical Hematology Atlas-E-Book*. Elsevier Health Sciences.

[Ruiz Segura, 2016] Ruiz Segura, F. (2016). *Algoritmos de visión computacional para la detección y clasificación de leucocitos en imágenes de frotis sanguíneos*. Tesis de Maestría, Universidad Autónoma de Aguascalientes, México.

[Saraswat and Arya, 2014] Saraswat, M. and Arya, K. V. (2014). Feature selection and classification of leukocytes using random forest. *Medical & biological engineering & computing*, 52(12):1041–1052.

[Sarrafzadeh et al., 2017] Sarrafzadeh, O., Dehnavi, A. M., Banaem, H. Y., Talebi, A., and Gharibi, A. (2017). The best texture features for leukocytes recognition. *Journal of medical signals and sensors*, 7(4):220.

[Sarrafzadeh et al., 2014] Sarrafzadeh, O., Rabbani, H., Talebi, A., and Banaem, H. U. (2014). Selection of the best features for leukocytes classification in blood smear microscopic images. In *Proceedings of the Medical Imaging 2014: Digital Pathology*, page 90410P.

[Shirazi et al., 2016] Shirazi, S. H., Umar, A. I., Naz, S., and Razzak, M. I. (2016). Efficient leukocyte segmentation and recognition in peripheral blood image. *Technology and Health Care*, 24(3):335–347.

[Standring, 2015] Standring, S. (2015). *Gray's Anatomy International Edition: The Anatomical Basis of Clinical Practice*. Elsevier Health Sciences.

[SyM, 2018] SyM (2018). Biometría hemática completa. Recuperado octubre de 2018, de: https://www.saludymedicinas.com.mx/centros-de-salud/embarazo/analisis-y-estudios-de-laboratorio/biometria-hematica-completa-hemograma.html.

[Tharwat, 2018] Tharwat, A. (2018). Classification assessment methods. *Applied Computing and Informatics*, pages 1–13.

[Theodoridis and Koutroumbas, 2008] Theodoridis, S. and Koutroumbas, K. (2008). *Pattern Recognition (Fourth Edition)*. Academic Press.

[Tuan Muda and Abdul Salam, 2013] Tuan Muda, T. Z. and Abdul Salam, R. (2013). Comparative analysis on blood cell image segmentation. In *Proceedings of the 2nd International Symposium on Computer, Communication, Control and Automation*, pages 474–477.

[Vapnik, 1998] Vapnik, V. (1998). *Statistical Learning Theory*. John Wiley & Sons.

[Wei and Cao, 2016] Wei, X. and Cao, Y. (2016). Automatic counting method for complex overlapping erythrocytes based on seed prediction in microscopic imaging. *Journal of Innovative Optical Health Sciences*, 9(5):1650016.

[Wei and Stocker, 2016] Wei, X.-X. and Stocker, A. A. (2016). Mutual information, fisher information, and efficient coding. *Neural computation*, 28(2):305–326.

[Wu et al., 2008] Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Philip, S. Y., et al. (2008). Top 10 algorithms in data mining. *Knowledge and information systems*, 14(1):1–37.

[Yen et al., 1995] Yen, J.-C., Chang, F.-J., and Chang, S. (1995). A new criterion for automatic multilevel thresholding. *IEEE Transactions on Image Processing*, 4(3):370–378.

[Zamani and Safabakhsh, 2006] Zamani, F. and Safabakhsh, R. (2006). An unsupervised GVF snake approach for white blood cell segmentation based on nucleus. In *Proceedings of the 8th International Conference on Signal Processing*, pages 1–4.

[Zayed and Elnemr, 2015] Zayed, N. and Elnemr, H. A. (2015). Statistical analysis of haralick texture features to discriminate lung abnormalities. *Journal of Biomedical Imaging*, 2015:1–12.

[Zhang et al., 2014] Zhang, C., Xiao, X., Li, X., Chen, Y.-J., Zhen, W., Chang, J., Zheng, C., and Liu, Z. (2014). White blood cell segmentation by color-space-based k-means clustering. *Sensors*, 14(9):16128–16147.

[Zhao et al., 2017] Zhao, J., Zhang, M., Zhou, Z., Chu, J., and Cao, F. (2017). Automatic detection and classification of leukocytes using convolutional neural networks. *Medical & biological engineering & computing*, 55(8):1287–1301.

[Zheng et al., 2018] Zheng, X., Wang, Y., Wang, G., and Liu, J. (2018). Fast and robust segmentation of white blood cell images by self-supervised learning. *Micron*, 107:55–71.

[Zhou, 2012] Zhou, Z.-H. (2012). *Ensemble methods: foundations and algorithms*. Chapman and Hall/CRC.

[Zuiderveld, 1994] Zuiderveld, K. (1994). Contrast limited adaptive histogram equalization. In Heckbert, P. S., editor, *Graphics Gems IV*, pages 474–485. Academic Press.

Anexo A

Células sanguíneas

A.1. Frotis sanguíneo

[Rodak and Carr, 2015] en su libro explican varios aspectos sobre el frotis sanguíneo. En primer lugar, el objetivo de la tinción de frotis de sangre periférica es identificar las células y reconocer con facilidad la morfología a través del microscopio. La tinción de Wright o de Wright-Giemsa es la utilizada con mayor frecuencia para los frotis de sangre periférica y de médula ósea. Estas tinciones contienen eosina y azul de metileno y, en consecuencia, se denominan tinciones policrómicas. Los colores presentan variaciones ligeras entre los diferentes laboratorios según el método de tinción.

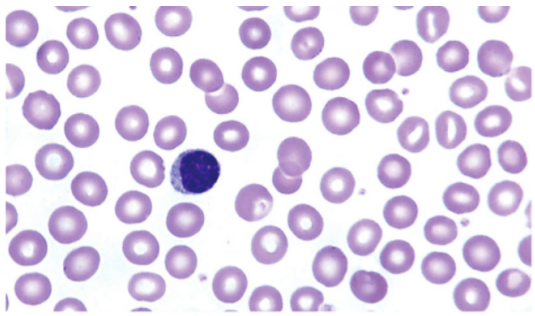


Figura A.1: Frotis de sangre periférica teñido de manera óptima que demuestra la zona adecuada para realizar la fórmula diferencial de leucocitos.

Las células se fijan en el portaobjetos por el metanol en la solución de tinción. Las reacciones de tinción dependen del pH, provocando múltiples colores en las células. Por ejemplo, los neutrófilos poseen gránulos citoplasmáticos que tienen pH neutro y aceptan algunas de las características de ambos colorantes. Un frotis teñido de manera óptima (ver Figura A.1) tiene las siguientes características:

1. Los eritrocitos deben ser de color rosado a salmón.
2. Los núcleos son de color azul oscuro o violeta.
3. Los gránulos citoplasmáticos del neutrófilo son de color lavanda a lila.
4. Los gránulos citoplasmáticos del basófilo son de color azul oscuro a negro.
5. Los gránulos citoplasmáticos de los eosinófilos son de color rojo o anaranjado.
6. La zona entre las células(fondo) debe ser incolora, clara, limpia y sin colorante precipitado.

Para la evaluación de la morfología celular es esencial realizar la preparación correcta del frotis de sangre. Aunque algunos analizadores automatizados preparan y tiñen los frotis de sangre de acuerdo con los criterios establecidos, en muchos lugares se sigue utilizando la preparación manual del frotis. En consecuencia, genera variaciones en las tinciones. [AL-Dulaimi et al., 2018] describe cada uno de los desafíos existentes al realizar la clasificación de los leucocitos de manera automática.

A.2. Leucocitos

Los glóbulos blancos o leucocitos son producidos en la médula ósea y se encuentran en la sangre y el sistema linfático. Un leucocito tiene uno o más núcleos. Esta característica ayuda a distinguirlos de otras células sanguíneas. La estructura del leucocito consiste de un núcleo, citoplasma y pared celular, como se muestra en la Figura A.2

Los leucocitos pertenecen al menos a cinco categorías diferentes como son: basófilos, eosinófilos, linfocitos, monocitos y neutrófilos, y se distinguen por su tamaño, forma nuclear e inclusiones citoplasmáticas. En la práctica, los leucocitos a menudo se dividen en dos grupos principales, esto es, los que tienen prominentes gránulos citoplásmicos, los granulocitos y los que no tienen. Por lo general, hay tres tipos de granulocitos presentes: neutrófilos, eosinófilos y basófilos.

Los neutrófilos son importantes en la defensa del cuerpo contra los microorganismos, suelen ser los leucocitos más abundantes, formando el $40-75\,\%$ de leucocitos presentes en las muestras de sangre. Estos tienen un diámetro de $12-14\ \mu m$ [Standring, 2015] y contienen un núcleo con dos a cinco lóbulos conectados por un fino hilo de cromatina. Los cayados (bandas) son neutrófilos inmaduros que aún no han completado su condensación nuclear y tienen un núcleo con forma de U. Los cayados reflejan una desviación hacia la izquierda en la maduración de los neutrófilos en un esfuerzo por hacer más células de manera más rápida. Los neutrófilos pueden proveer claves para una variedad de alteraciones [Longo, 2012]. La Figura A.2 muestra un neutrófilo en banda y otro segmentado.

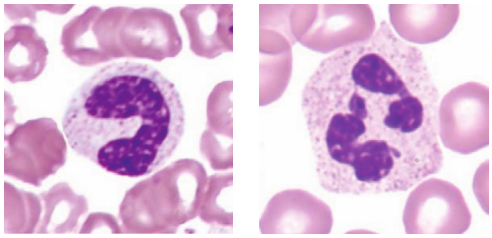


Figura A.2: Neutrófilo a) banda y b) segmentado.

Los basófilos (ver Figura A.3) son más pequeños que otros granulocitos, su tamaño es de $10-14\ \mu m$ en diámetro, y son aún más escasos que los eosinófilos en la sangre. Estos forman sólo $0.5-1\,\%$ de la población leucocitaria total de sangre normal [Standring, 2015]. Su característica distintiva es la presencia de gránulos grandes y conspicuos. El núcleo es algo irregular o bilobulado, y generalmente está oculto en frotis de sangre manchados por el color similar de los gránulos basófilos.

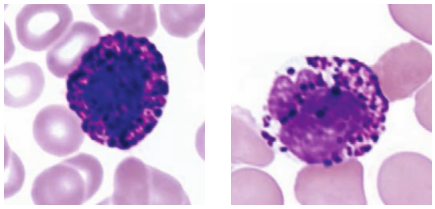


Figura A.3: Basófilos.

Los eosinófilos se reconocen fácilmente en frotis teñidos por sus grandes gránulos (ver Figura A.4). Sus gránulos específicos citoplasmáticos son uniformemente grandes de 0.5 μm. El núcleo tiene dos prominentes lóbulos conectados por un fino hilo de cromatina. El diámetro de la célula suele oscilar entre $12-15$ μm y representan entre $1-4\,\%$ de la población leucocitaria. Los eosinófilos son ligeramente más grandes que los neutrófilos, muestran núcleos bilobulados y contienen grandes gránulos rojos. Las enfermedades de los eosinófilos están vinculadas con demasiados de ellos, más que con cambios morfológicos o cualitativos [Longo, 2012].

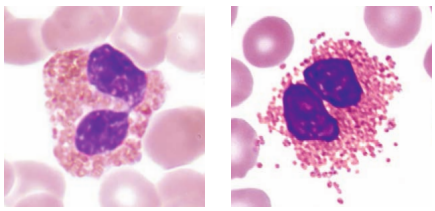


Figura A.4: Eosinófilos.

A continuación se describen los leucocitos que no tienen granulocitos en su citoplasma y son conocidos como mononucleares debido a que sólo tienen un núcleo (ver Figura A.5). Los linfocitos son el segundo tipo de célula más numeroso en un adulto, formando un $20-30\,\%$ de la población leucocitaria. La mayoría de los linfocitos circulantes son pequeños, de 6 a 8 μm de diámetro; algunos son de tamaño mediano y tienen un volumen citoplásmico aumentado, a menudo en respuesta a la estimulación antigénica. En ocasiones, se observan células de hasta 16 μm en la sangre periférica [Standring, 2015].

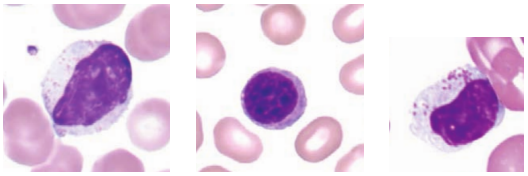
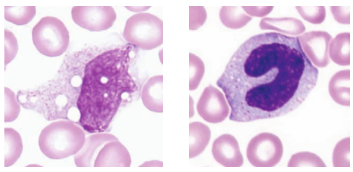


Figura A.5: Linfocitos.

Los monocitos son los más grandes en tamaño de los leucocitos (15 a 20 μm de diámetro), pero forman solo una pequeña proporción de la población total. Su núcleo, que es eucromático, es relativamente grande e irregular, pero por lo general aparece doblado [Standring, 2015]. El citoplasma es de coloración pálida, particulado y típicamente vacuolado (ver Figura A.6) .



(a) (b)

Figura A.6: Monocito con a) vacuolas y b) sin vacuolas.

A.3. Eritrocitos

El eritrocito o glóbulo rojo mide aproximadamente 7 μm de diámetro, por su forma bicóncava, se aprecia una palidez central que corresponde a una tercera parte de su diámetro. Una forma sencilla de valorar su tamaño es compararlo con el núcleo del linfocito, que en condiciones normales es casi de la misma dimensión. Aparecen como las células más abundantes en el frotis de sangre periférica. Los eritrocitos tienen una membrana plasmática que encierra principalmente una proteína única, la hemoglobina, como una solución al 33 %. La Figura A.1 muestra eritrocitos maduros en un frotis.

Anexo B

Características seleccionadas para el reconocimiento de leucocitos

El mejor modelo de clasificación de leucocitos entrenado para el reconocimiento de leucocitos y descrito en la Subsección 4.3.3 utiliza 49 caracteracterísticas, las cuales fueron obtenidas mediante el método de selección de características basada en un umbral (TFS). A continuación se muestran las 49 características seleccionadas en el Cuadro B.2, donde la abreviación de dichas características es mostrada en el Cuadro B.1.

Para el caso de clasificación usando como clases tres tipos de leucocitos (granulocitos, linfocitos y monocitos), al aplicar los métodos de selección de características se obtuvo el mejor modelo usando 49 características. Éstas fueron seleccionadas utilizando el método selección de características univariadas (UFS). El Cuadro B.3 lista esas características.

El enfoque de clasificación de leucocitos por subtipos obtiene su mejor modelo utilizando 109 características, las cuales fueron seleccionadas mediante el método de selección de características basada en un umbral (TFS). Las características seleccionadas se listan en el Cuadro B.4.

Característica	**Notación**
Área	Area
Perímetro	Perimetro
Compacidad	compacidad
Dispersión	IR
7 Momentos de Hu	mHu1 ..... mHu7
14 características de textura de Haralick	Ha1 .....Ha14
Momentos de Zernike: Grado 8	momZer_Degre8_1 . . momZer_Degre8_25
Media y varianza de cada componente del espacio de color RGB	R_mean, R_var G_mean, G_var B_mean, B_var
Media y varianza de cada componente del espacio de color HSV	H_mean, H_var S_mean, S_var V_mean, V_var
Media y varianza de cada componente del espacio de color YCbCr	Y_mean, Y_var Cb_mean, Cb_var Cr_mean, Cr_var
Media y varianza de cada componente del espacio de color L*a*b.	L_mean, L_var a_mean, a_var b_mean, b_var

Cuadro B.1: Notación de las características utlizadas.

	Núcleo	Citoplasma	Célula
Características	Area Perimetro compacidad mhu1 Ha1 Ha3 Ha6 Ha8 Ha14 R_mean G_mean G_var B_mean H_mean S_mean S_var V_mean Y_mean Cb_mean Cb_var Cr_mean L_mean a_mean a_var b_mean	Area Perimetro mhu1 Ha1 Ha5 Ha6 Ha8 Ha9 Ha11 monZer_Degre8_2 R_mean R_var G_mean H_mean S_mean V_mean Y_mean Cb_mean Cr_mean L_mean L_var a_mean	Ha3 Ha14

Cuadro B.2: Características seleccionadas para el enfoque de clasificación con cinco clases.

	Núcleo	**Citoplasma**	**Célula**
Características	Area compacidad IR mhu3 Ha1 Ha3 Ha8 Ha13 Ha14 R_mean H_mean V_mean Y_mean Cr_mean L_mean b_mean	Area mhu1 mhu3 mhu4 mhu5 mhu6 mhu7 Ha1 Ha5 Ha6 Ha8 Ha9 Ha10 Ha11 monZer_Degre8_2 R_mean R_var G_mean B_mean H_mean H_var S_mean V_mean Y_mean Y_var Cb_mean Cr_mean L_mean L_var a_mean b_mean	Ha12 Ha14

Cuadro B.3: Características seleccionadas para el enfoque de clasificación de leucocitosen dos etapas: tipo.

	Núcleo	**Citoplasma**	**Célula**
Características	Perimetro compacidad mhu1 mhu2 mhu4 mhu5 mhu6 mhu7 Ha1 Ha3 Ha4 Ha6 Ha7 Ha10 Ha14 R_var G_mean G_var B_var H_var S_mean S_var V_var Y_var Cb_mean Cb_var Cr_var L_var a_mean a_var b_var	Perimetro compacidad mhu1, mhu4 mhu7 , Ha1 Ha2 , Ha3 Ha4 , Ha6 Ha7 , Ha9 Ha13 , Ha14 monZer_Degre8_2 monZer_Degre8_4 monZer_Degre8_6 monZer_Degre8_8 monZer_Degre8_13 monZer_Degre8_25 R_mean R_var G_mean G_var B_mean B_var H_mean H_var S_mean S_var V_mean V_var Y_mean Y_var Cb_mean Cb_var Cr_mean Cr_var L_mean L_var a_mean a_var b_mean b_var	mhu1 mhu2 mhu6 Ha2 Ha3 Ha4 Ha6 Ha7 Ha8 Ha9 Ha10 Ha11 Ha12 Ha14 R_mean R_var G_mean G_var B_mean H_mean H_var S_mean S_var V_mean Y_mean Y_var Cb_mean Cb_var Cr_mean L_mean L_var a_mean a_var b_mean

Cuadro B.4: Características selecionadas para el enfoque de clasificación de leucocitos en dos etapas: subtipos.

Anexo C

Configuración y manual de usuario de la aplicación Web

La aplicación desarrollada en la Sección 3.3 requiere dos componentes importantes, los cuales son el API-REST y la aplicación Web. Para replicar este proyecto, es necesario contar con la configuración y el software adecuado. En este anexo se explican la configuración y se presenta un manual de usuario de la aplicación.

C.1. Configuración

Angular fue utilizado para crear el frontend de la aplicación, entonces se debe tener instalado el software Node.js y NPM (*Node Package Manager*). Estas herramientas se instalan simultáneamente al descargar y ejecutar el instalador oficial de la página de Node.js [Nodejs, 2019]. Cuando la instalación haya finalizado, en la terminal se debe escribir el siguiente comando `node -v` y si la instalación fue exitosa, la versión de node será mostrada.

Cualquiera que sea el método de instalación, debe tener la versión más reciente de Node.js (es decir, $>= 8$). Al instalar Nodejs, automáticamente se instala NPM, el cual es un manejador de paquetes que es utilizado para instalar, compartir y distribuir código, en este caso se utiliza el comando npm para instalar Angular-CLI. Para iniciar el frontend desde la interfaz de línea de comandos (CLI) es necesario ejecutar el siguiente comando en una consola:

```
npm install -g @angular/cli
```

A continuación, se describe el proceso de configuración del *backend* de la aplicación. Para esto es necesario contar con PIP, el cual es un sistema de gestión de paquetes utilizado para instalar y administrar paquetes de software escritos en Python. La instalación de PIP se muestra a continuación:

```
sudo apt-get update && sudo apt-get -y upgrade
sudo apt-get install python-pip
```

Antes de ejecutar el proyecto, es necesario instalar las bibliotecas requeridas. Para esto, se debe cambiar a la carpeta `Backend` del proyecto que aloja el código de la API-REST desarrollada. Posteriormente, una vez en el directorio se instalan las bibliotecas para el correcto funcionamiento. Para ello escribir en la terminal lo siguiente:

```
/AI_BLOODAPP/Backend$ pipenv install -r requirements.txt
```

El archivo requirements.txt contiene todas las bibliotecas a instalar. Hasta este punto, se tienen instaladas todas las herramientas y bibliotecas necesarias. Para iniciar toda la aplicación se requiere ejecutar 2 scripts en distintas terminales. Éstos se describen a continuación:

- API-REST. Se debe posicionar en la carpeta raíz del proyecto y escribir en la terminal `./backend.sh &` para ejecutar la API. En la Figura C.1 se muestra la ejecución de este script.
- La aplicación Web. Se debe posicionar en la carpeta raíz del proyecto y escribir en la terminal `./APP.sh &` para ejecutar la aplicación Web. En la Figura C.2 se muestra la ejecución de este script.

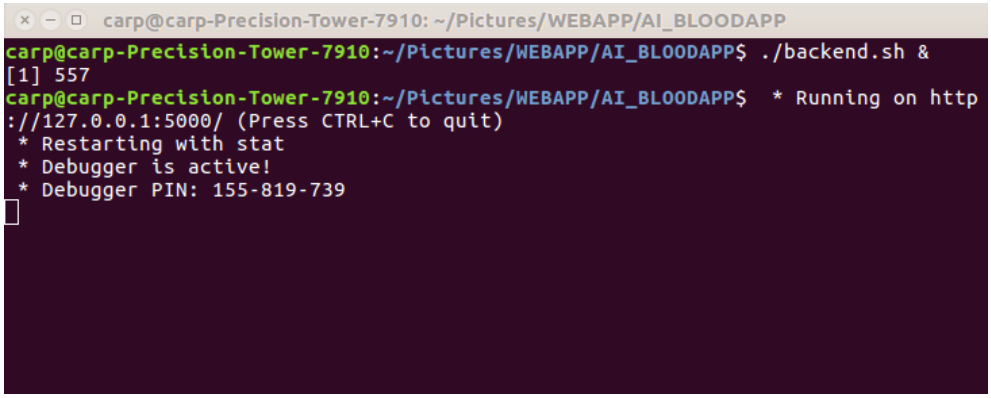


Figura C.1: Ejemplo de cómo iniciar la API-REST.

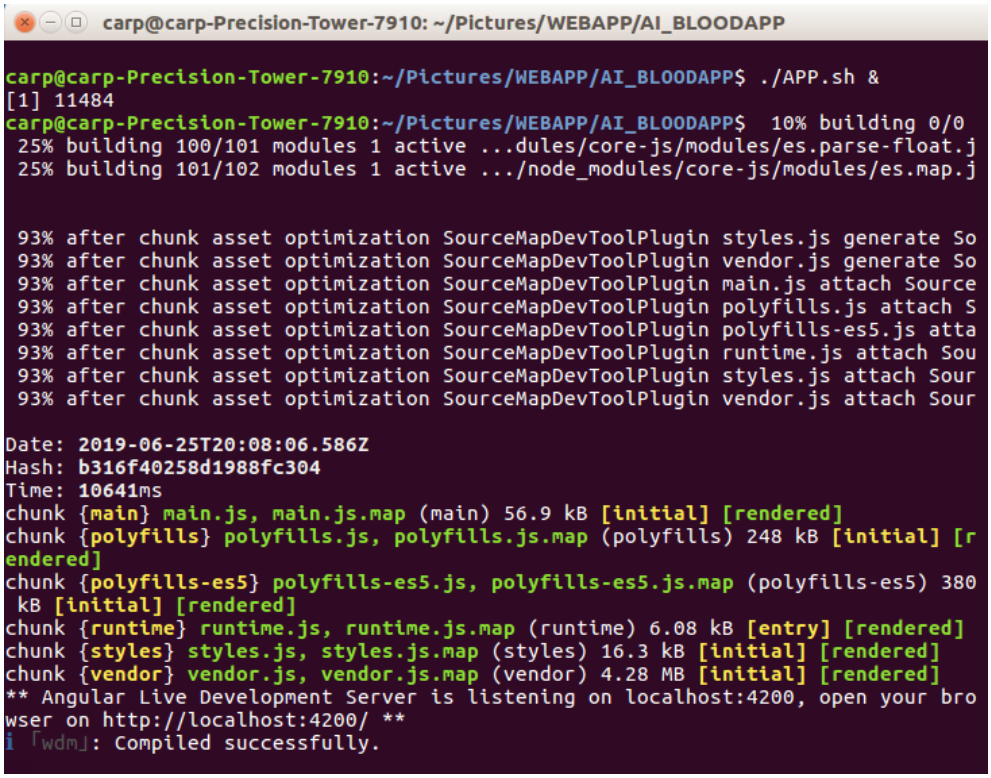


Figura C.2: Ejemplo de cómo iniciar la aplicación Web.

C.2. Manual de usuario

La aplicación web BloodCell.Ai es un software que permite identificar y contar leucocitos y eritrocitos, además, de reconocer cinco tipos de leucocitos.

Para hacer uso de la aplicación web es necesario acceder a través de la IP y apuntando al puerto 4200. Esto es, escribir en la barra del navegador Web la dirección *http://localhost:4200/* (ver Figura C.3). Al iniciar, se muestra la pantalla principal de la aplicación, en la parte superior se visualizan dos opciones. La primera opción es "Inicio" la cual sirve para dirigirse a la pantalla principal de la aplicación. La segunda opción es "Procesar frotis sanguíneo", la cual sirve para cargar la imagen de frotis (ver Figura C.4).

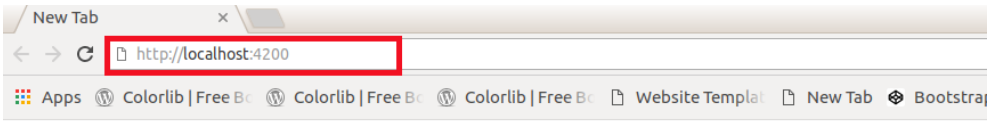


Figura C.3: Ingresar a la aplicación desde el navegador Web.

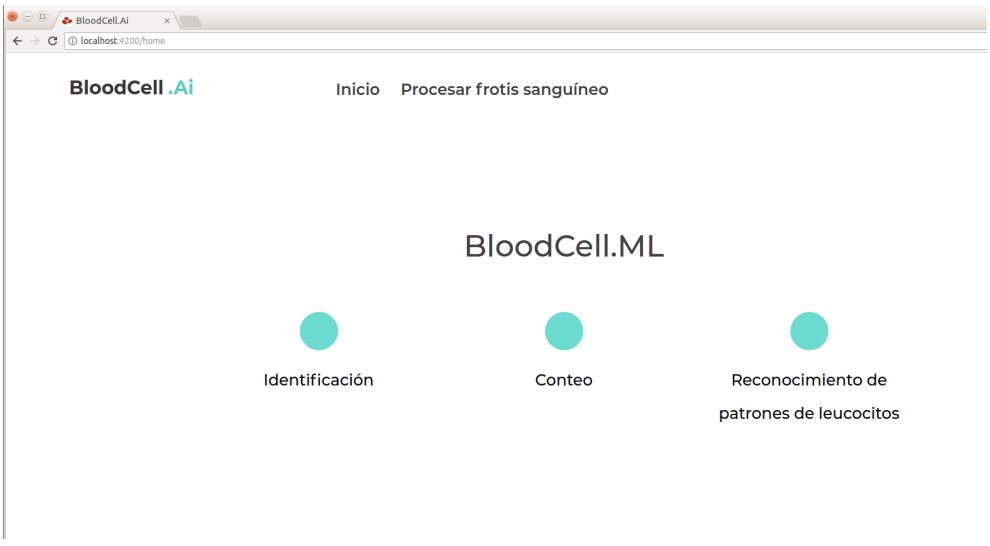


Figura C.4: Página principal de la aplicación.

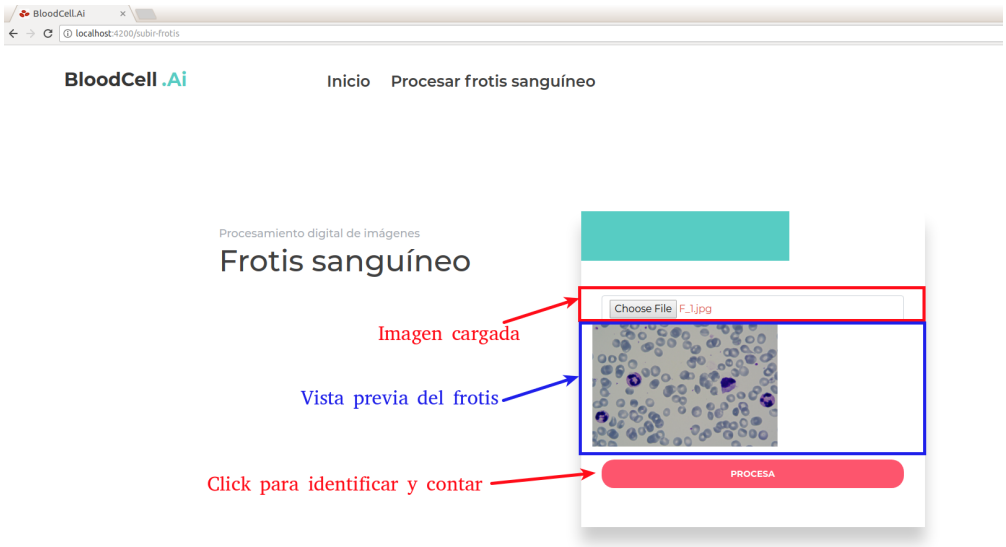
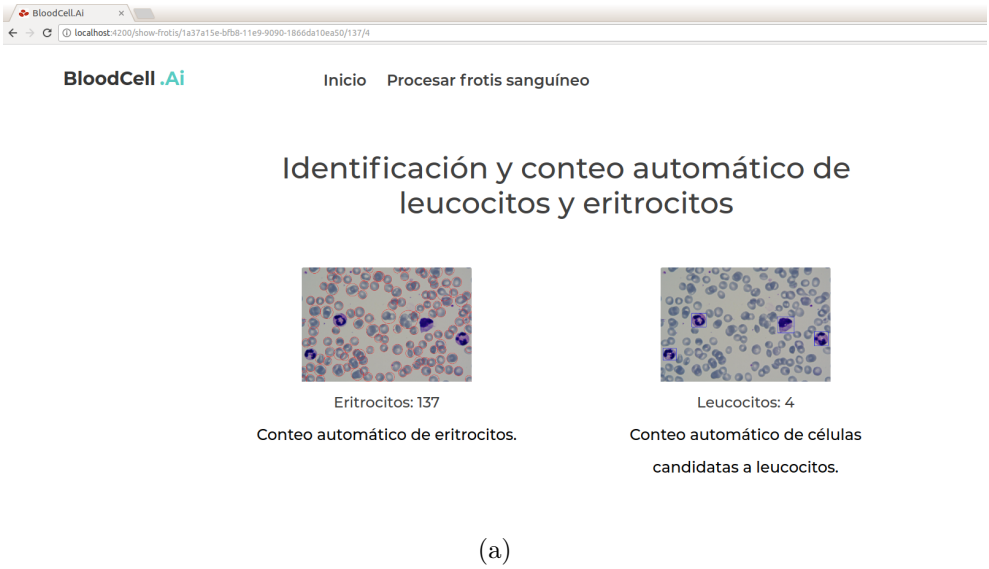


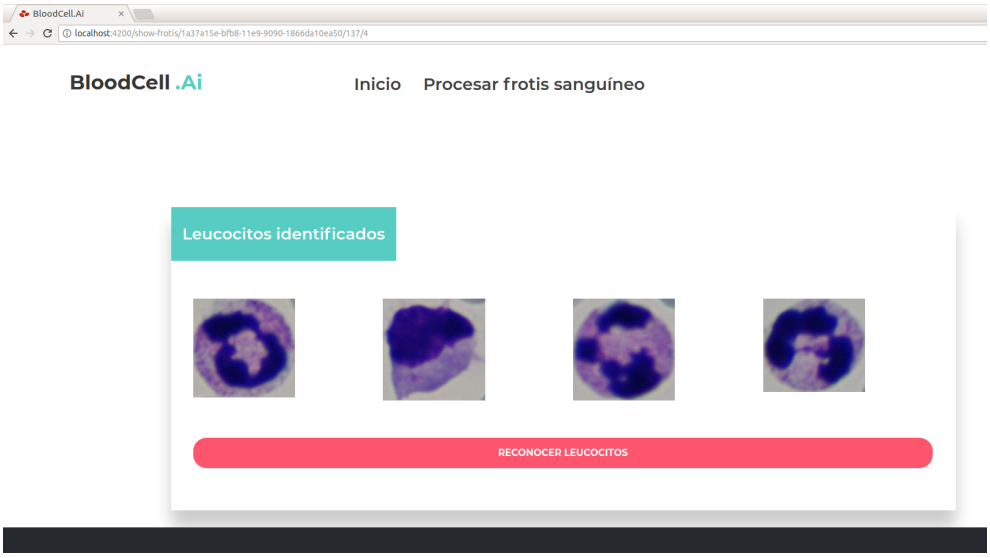
Figura C.5: Página para subir la imagen de frotis sanguíneo.

Para cargar una imagen se tiene que dar clic al botón “Choose file” y seleccionar la imagen a procesar. Inmediatamente al cargar la imagen, una previsualización de ésta es mostrada. Una vez que se ha elegido la imagen dar clic en el botón “Procesar” (ver Figura C.5), lo cual inicia la identificación y conteo de leucocitos y eritrocitos. La Figura C.6 ilustra un ejemplo de los resultados de identificación y conteo al finalizar este proceso. Se puede observar en la parte superior los resultados de conteo de leucocitos y eritrocitos, y en la parte inferior los leucocitos identificados. Si el usuario desea ver las imágenes en mayor tamaño, solo debe dar clic en alguna de éstas.

A continuación, para realizar el reconocimiento de los leucocitos identificados, dar clic en el botón “Reconocer”. La aplicación redirige a una pantalla para esperar los resultados (ver Figura C.7). Una vez finalizado el reconocimiento de patrones de leucocitos se muestran los resultados obtenidos para los 4 leucocitos identificados en la etapa anterior (ver Figura C.8).



(a)



(b)

Figura C.6: Resultados: a) conteo de eritrocitos y leucocitos, y b) leucocitos identificados.

Figura C.7: Página de espera.

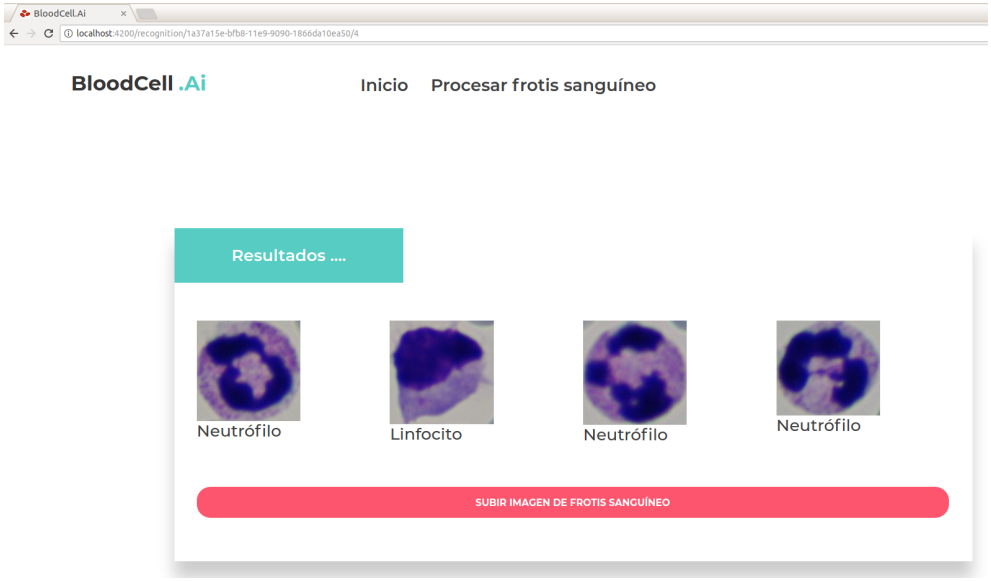


Figura C.8: Página de resultados de leucocitos.