Detección de blastos de leucemia mieloide aguda en extendido de sangre periférica por medio de técnicas de aprendizaje automático.

Arley Stiven Quintero Espinal

Trabajo de grado como requisito para optar al título de Ingeniero de Telecomunicaciones

Tutor
Jhon James Granada Torres - PhD

Universidad de Antioquia
Facultad de Ingeniería, Departamento de Ingeniería electrónica y telecomunicaciones
Pregrado UdeA
Medellín
2022

Cita	(Quintero Espinal Arley Stiven, 2022) [1]
Referencia Estilo IEEE (2022)	[1] Quintero Espinal Arley Stiven. (2022). [Trabajo de grado]. Universidad de Antioquia, Medellín.

creative commons BY NC SA

Universidad de Antioquia - www.udea.edu.co

Rector: John Jairo Arboleda Cespedes.

Decano/Director: Jesús Franciso Vargas Bonilla.

Jefe departamento: Augusto Enrique Salazar Jimenez.

El contenido de esta obra corresponde al derecho de expresión de los autores y no compromete el pensamiento institucional de la Universidad de Antioquia ni desata su responsabilidad frente a terceros. Los autores asumen la responsabilidad por los derechos de autor y conexos.

Resumen

La detección de leucemia mieloide aguda se realiza a través de pruebas morfológicas de sangre periférica como método de diagnóstico rápido. Este análisis consiste en contabilizar y diferenciar morfológicamente glóbulos blancos y rojos, las plaquetas y realizar la medición de la hemoglobina de la sangre. Luego de realizados estos procesos se determina si los diferentes tipos de células son anormales.

Generalmente el proceso de reconocimiento de células cancerígenas se realiza de manera visual en un microscopio. Sin embargo, requiere en ocasiones de ayuda tecnológica, sobre todo en casos donde la probabilidad de confusión de patrones es alta, ejemplo de esto, el análisis de extendido periférico donde se obtiene una gran cantidad de distintos tipos de célula.

El apoyo para realizar diagnósticos de detección de leucemia utilizando herramientas digitales ha aumentado, por lo cual se hace indispensable la clasificación de las variantes de la leucemia aguda, esto con el fin de ofrecer un tratamiento adecuado que permita mejorar las condiciones de vida.

Por esta razón, en áreas como la ingeniería, se proponen métodos de reconocimiento de patrones que permitan realizar la clasificación de células cuyo objetivo es mejorar y/o apoyar el diagnóstico de enfermedades.

De acuerdo con lo anterior, en este trabajo se propone la detección de leucemia aguda en imágenes de extendidos de sangre periférica de pacientes, utilizando un sistema de aprendizaje automático comparando diferentes algoritmos de aprendizaje automático.

1. Introducción

La leucemia mieloide aguda (*LMA*) es una enfermedad maligna que presenta anormalidades genéticas asociadas a reordenamientos cromosómicos estructurales. La mayoría de estos reordenamientos generan distintos grupos de enfermedades cuyas características morfológicas e inmunofenotípicas difieren una de otras [1]. De éstas se encuentran alrededor de 13 mutaciones por tipo de *LMA*, lo que hace difícil el proceso de identificación. A nivel clínico se realiza la detección a través de las señales físicas características de la mutación, pero en algunos casos se recurre al análisis de médula ósea para determinar la cantidad de células blásticas [2].

Para realizar análisis e identificación de leucemia se ha utilizado la microscopía ya que arroja resultados morfológicos importantes. Las características morfológicas comunes incluyen la presencia de abundante citoplasma basofílico, que a menudo muestran gránulos muy grandes. De esta forma los principales diagnósticos se centran en el conteo cuidadoso de células blásticas, el cumplimiento de los criterios de diagnóstico para la displasia morfológica y la evaluación de las anomalías citogenéticas [3].

En campos de estudio como son el aprendizaje automático (Del inglés: *Machine Learning*) y el Aprendizaje Profundo (Del inglés *Deep Learning*) se mejora la predicción debido a la extracción de características morfológicas. En métodos de aprendizaje automático de agrupación (Del inglés: *K-Means*), histograma de ecualización, *SVM*, se realiza un pre-procesamiento, posteriormente una segmentación y una extracción de características. Entre los resultados de cada clasificador se pueden encontrar mejoras en la segmentación, buenos porcentajes de acierto en la detección de células, así como la separación de células que se traslapan. Además de ser rápidos en su ejecución permiten obtener buenos resultados de clasificación [4].

Se evidencia la necesidad de implementar un método automatizado para el reconocimiento y clasificación de leucemia en extendido de sangre periférica. Sin embargo, la clasificación de los tipos de leucemia por medio de la morfología de la célula en imágenes digitales es un aspecto que nuestro medio aún no ha sido abordado a profundidad y se hace indispensable la clasificación de las variantes de la leucemia aguda con el fin de ofrecer un tratamiento adecuado que permita mejorar las condiciones de vida de los pacientes.

Por lo anterior en este trabajo se elabora un método que permite realizar la segmentación de células cancerígenas en los siguientes conjuntos de entrenamiento: eosinófilos, monocitos, neutrófilos, linfocitos y las de control eosinófilos, monocitos, neutrófilos y

linfocitos sanos, con base en la extracción de características morfológicas como lo son la excentricidad, redondez y solidez. Una vez se obtiene el conjunto de datos para la matriz de entrenamiento y la matriz de pruebas se aplica reducción de dimensiones (Del inglés: *Principal Component Analysis- PCA*).

Los algoritmos de clasificación que se utilizan son vecino más cercano (Del inglés: *K-Nearest Neighbors*), máquinas de soporte vectorial lineal (Del inglés: *Linear Vector Support Machines*) y bosque aleatorio o bosque al azar (Del inglés: *Random Forest)* cada uno probados de dos formas: la primera, utilizando reducción de dimensiones y la segunda prueba sin incluirlas.

Los escenarios propuestos de prueba para los algoritmos bajo la restricción del uso y el no uso de PCA son los siguientes: pruebas de clasificación con dos conjuntos de entrenamiento, donde se comparan muestras del mismo tipo de célula diferenciando células cancerígenas de células sanas.

Por último, se valida el funcionamiento de los algoritmos con base en los porcentajes de acierto según los parámetros exactitud, precisión y exactitud.

El desarrollo general de este proyecto se compone de secciones donde en cada una se evidencian los pasos a desarrollar para cada una de las etapas. De la siguiente manera: En la sección 2 se presentan los objetivos tanto general como específicos para el desarrollo de la metodología y posterior análisis de resultados. En la sección 3 se presenta el marco teórico y estado del arte. Aquí se da un vistazo general del proceso de detección de leucemia desde el punto de vista clínico, las técnicas para su detección vistas desde la medicina, como desde el análisis de imágenes digitales por medio de algoritmos de clasificación.

En la sección 4 se presenta la metodología con cada uno de los pasos realizados desde la etapa de pre-procesamiento hasta la etapa de clasificación. Los resultados de cada uno de los pasos realizados se presentan en la sección 5; aquí se muestran los escenarios planteados y los valores óptimos obtenidos del proceso de clasificación y la discusión de los resultados obtenidos.

Para el final, en la sección 6 se presentan las conclusiones del proceso realizado, los resultados obtenidos y sugerencias para trabajos futuros.

2. Objetivo General

Detectar blastos de leucemia mieloide aguda en extendido de sangre periférica por medio de técnicas de aprendizaje automático.

2.1 Objetivos específicos.

- Obtener una base de datos de imágenes de extendido de sangre periférica de pacientes con leucemia mieloide aguda y sujetos sanos para comparar los blastos.

- Aplicar técnicas de procesamiento digital de imágenes utilizadas como entrada en un clasificador basado en aprendizaje automático.

- Diseñar e implementar en Software un algoritmo de aprendizaje automático y que permita detectar blastos de leucemia mieloide aguda.

- Validar el desempeño de los algoritmos diseñados y comparar las tasas de acierto entre pacientes con leucemias e individuos sanos.

3. Marco Teórico y estado del arte

3.1 La leucemia mieloide aguda y tipos de la enfermedad. Es una enfermedad a nivel de células blancas que presenta anormalidades genéticas asociadas a reordenamientos cromosómicos estructurales. Inicialmente se proliferan en la médula ósea antes de propagarse a la sangre periférica, bazo, ganglios linfáticos y finalmente a los demás tejidos. La principal característica de este tipo de células es un defecto en la maduración en su fase de desarrollo [5]. La mayoría de estos reordenamientos generan distintos tipos de enfermedades las cuales difieren de características morfológicas. De éstas se encuentran un gran número de mutaciones generando alrededor de 13 tipos de *LMA*, lo que hace difícil el proceso de identificación [6]. Algunos tipos de leucemia son las siguientes:

- Leucemia promielocítica aguda en la cual predominan promielocitos anormales con forma de riñón. Algunas características morfológicas que caracterizan a este tipo de leucemia se presentan como aparente ausencia de gránulos y núcleos bilobulados.

- Leucemia mieloide aguda (megacarioblástica) la cual genera anemia y un recuento moderado de glóbulos blancos.

Otros tipos de leucemia involucran eosinófilos inmaduros, monocitos y linfocitos cuyas características clínicas van desde sarcomas extramedulares, filtración de tejidos como encías, piel, entre otros.

3.2 Técnicas para el diagnóstico de leucemia mieloide aguda. Cada tipo leucemia presenta características distintas de la enfermedad y, por lo tanto, se tienen diferentes opciones para el tratamiento. El examen físico inicia con un estudio del cuerpo para comprobar los signos vitales de salud, incluyendo control de síntomas de la enfermedad, como la aparición de masas o cualquier otro síntoma físico anormal [7]. Con base en la examinación de células sanguíneas existen diagnósticos para su detección que involucran criterios de citoquímica, inmunología, citogenética y biología molecular. En general, el inmunofenotipo representa una herramienta útil, ya que permite la detección de diferentes antígenos que identifican las etapas de maduración de las células afectadas [8].
Otros métodos incluyen translocaciones cromosómicas las cuales asocian más de 80 puntos de ruptura y variaciones de ensamblaje de ARN. Las muestras sintetizadas en

PCR se analizan a través de electroforesis en un gel de agarosa al 1.5%; luego en una banda de control se verifica la integridad del ARN [9].
Un recuento sanguíneo completo es una prueba para contabilizar el número glóbulos blancos en una muestra de sangre periférica o de médula ósea donde se mide su tamaño y madurez en los diferentes tipos de células sanguíneas [10].
El diagnóstico morfológico de la sangre consiste en contabilizar los glóbulos blancos y medir la hemoglobina de la sangre. Luego se examinan morfológicamente las diferentes células y se determinan si son anormales; este procedimiento en Colombia es realizado por un profesional de microbiología y bioanálisis especializado en hematología con experiencia en la detección de la enfermedad. El problema con este tipo de análisis es que es lento y no es tan preciso debido a la subjetividad, ya que depende de las capacidades y experiencia del hematólogo [11]. Por otro lado, se proponen diagnósticos a través de métodos automatizados para el reconocimiento y clasificación de subtipos de leucemia aguda en imágenes digitales de células sanguíneas por medio de la morfología.

3.3 Algoritmos de clasificación. Los problemas de clasificación son instancias del aprendizaje supervisado donde se dispone de un conjunto de entrenamiento de observaciones previamente identificadas. En general, se pretende crear un modelo que aprenda de estos datos etiquetados para que pueda predecir futuros datos sin etiquetas [12]. Actualmente existen algoritmos de clasificación entre ellos a) *SVM*, b) *K-NN*, c) Bosque aleatorio. En el primero se encuentran las máquinas de soporte vectorial de tipo Gaussiano el cual separa los puntos de datos que pertenecen a diferentes clases con un límite de decisión. Sus objetivos son aumentar la distancia del límite de decisión a las clases y maximizar la cantidad de puntos que se clasifican correctamente en el conjunto de entrenamiento [13]. Otro tipo de clasificador dentro de las *SVM* es el de tipo lineal. Este se utiliza cuando los datos son linealmente separables y cuando hay una gran cantidad de funciones en un conjunto de datos.
El *algoritmo K-NN* clasifica por medio de la escogencia del vecino más cercano. Según sea la elección los objetos de la imagen se asignan a las clases relevantes. En el caso de leucemia aguda, el algoritmo puede ser utilizado para obtener un resultado de clasificación determinando cuáles células son normales o anormales [14].
Para el caso del clasificador bosque aleatorio, se crea un árbol de decisión partiendo de una secuencia aleatoria de un subconjunto de datos del conjunto de entrenamiento. La clasificación es realizada de la siguiente forma: dado un ejemplo con sus respectivas características se sitúa en cada uno de los árboles de decisión que hay en el bosque y el que el algoritmo ha creado. Cada árbol realizará una clasificación y ese árbol dará una votación a la clase que ha predicho [15].

3.4 Utilización de técnicas de procesamiento digital de imágenes. Varios estudios han utilizado métodos para la segmentación y clasificación de células sanguíneas en

sangre periférica utilizando extracción de características de tipo morfológico. El estudio de *Kan Jiang et al* [16], presenta un método de segmentación de glóbulos blancos que utiliza un filtro espacio-escalar. Primero, se segmenta un glóbulo blanco generando una sub-imagen, luego se utiliza el filtro espacio-escalar para extraer el núcleo del glóbulo blanco. Posteriormente se agrupa mediante línea divisoria de aguas (Del inglés: *Watershed*) utilizando el histograma del espacio de color *HSV* (Del inglés: *Hue, Saturation, Value*) para extraer el citoplasma del glóbulo blanco y por último se realizan operaciones morfológicas (post-procesamiento) para obtener la conexión completa del glóbulo blanco. En el estudio se procesaron 45 imágenes y se obtuvieron 50 glóbulos blancos, la exactitud de este paso fue del 100%, ya que pudieron reconocer los 50 glóbulos blancos en las imágenes.

En otro estudio se realizó el *Análisis de imágenes biomédicas en color procedente de microscopía* [17]; en este se realizó una estrategia de segmentación en tres etapas: a) selección del espacio de color de representación de los colores, b) técnicas de procesamiento vectorial empleando morfología en color, c) proceso de segmentación en dos etapas, una de determinación de las regiones de interés y la segunda de refinamiento en la segmentación de los cuerpos biológicos. Según los pasos establecidos se determinan las regiones de interés. Lo relevante de esta solución fue la idea de hacer pre-procesamiento a las imágenes de leucemia antes de empezar con la segmentación.

También se han propuesto métodos que utilizan una etapa de filtrado [*W. Srisukkham, et al* 18]. En el estudio en mención se realizó la detección de leucemia aguda por fases. En la fase de preprocesamiento se eliminó el ruido a través de la técnica de filtrado medio seguido de la técnica de enmascaramiento y luego se convirtieron las imágenes de sangre del espacio de color rojo, verde y azul (Del inglés: *Red, Green, Blue*) a *HSV*. Se segmentó utilizando el método Otsu (umbralización). Las características de forma y textura tanto de células normales como anormales se clasificaron mediante la máquina de soporte vectorial *SVM.*

Otros estudios han realizado comparaciones entre muestras de sangre normal y de personas con leucemia utilizando técnicas de procesamiento de imágenes como *K*-Means, transformada de línea divisoria, ecualización de histograma y características basadas en la forma. En uso de estos, un aspecto importante en la técnica de extracción de características utilizada es que ayudó en la detección de células blancas superpuestas obteniendo una precisión del 97,8% [*Dharani, & Hariprasath S* 19].

El sistema propuesto en *Ahmed et al* [20], empieza con la segmentación de las células blancas. En este se incluyó la conversión de *RGB* al modelo *CMYK* (Del inglés: *Cyan, Magenta, Yellow, Key*), histograma de ecualización por técnica Zach y operación de remoción de fondo. Las características extraídas incluyen, color, textura y características de la forma. Luego cada característica fue expuesta a tres técnicas de normalización *z-score*, *min-max*, y escala de grises para acortar la distancia entre los valores de las características. Luego de aplicar diferentes clasificadores se llegó a la conclusión de que el clasificador K-NN, presenta mayor precisión. El algoritmo propuesto se ejecutó en 100 imágenes microscópicas de células malignas y el resultado experimental muestra que logró una buena segmentación de éstas a partir de su complicado entorno. Para este caso la precisión alcanzada fue de 98.38%.

4. Metodología

4.1 Obtención de una base de datos de imágenes. La información utilizada para el desarrollo del proceso de segmentación es una base de datos libre, ya que debido a la contingencia mundial producida por la pandemia no se pudieron tomar muestras con apoyo de la IPS universitaria. El grupo posee células enfermas de tipo: eosinófilos, monocitos, neutrófilos y linfocitos. El donante de las células correspondientes a pacientes sanos solicitó no proporcionar la identidad.
Cada conjunto de células posee un total de 80 muestras, dando un total para la base de datos de 640 imágenes.
Algunos aspectos que retrasaron el desarrollo del código para la clasificación se mencionan a continuación.

En primer lugar, muchas de las librerías utilizadas en distintos estudios no se encuentran disponibles para la versión de jupyter con la que se trabajó, por lo cual se fue cambiando la perspectiva en cuanto a la metodología empleada para el desarrollo de las actividades planeadas.

4.2 Aplicación de las técnicas de procesamiento digital de imágenes. Como paso inicial, se importaron las librerías que permiten el funcionamiento del código. Por medio de la biblioteca *Scikit-Learn* de Python se aprovechan las utilidades para el análisis de datos y uso de algoritmos de clasificación.
Una vez almacenadas las imágenes se realizó el *pre-procesamiento*, es decir, se utilizaron las técnicas de segmentación y reconocimiento de patrones
Se aplicó *ecualización de histograma a los conjuntos de prueba de la base de datos*. Este método ayuda a obtener una distribución uniforme de los distintos niveles de intensidad, lo que permite mejorar el contraste en las imágenes [21].
El siguiente paso consistió en *utilizar el espacio de colores HSV a las imágenes ecualizadas*. Con este procedimiento se separaron cada uno de los canales que conforman el espacio para determinar cuál brindaba mejor visualización. Se optó entonces por utilizar el canal *S*, el cual hace referencia a la pureza del color [22].
Una vez extraído el canal *S* se *utilizó el filtro de media con adición de ruido sal y pimienta*. Este filtro permite mejorar las imágenes mediante el suavizado y la eliminación de ruidos, manteniendo la información de borde [23], que para este caso se ajusta para la extracción de características morfológicas.
Luego de filtrada la imagen se *utilizó la detección de bordes y marcación del contorno*. A través de la función de umbralización (Del inglés: *Tresholding*) se asignan valores de píxel en relación con el valor de umbral proporcionado. Es decir, si el valor del píxel es menor que el umbral, este queda con valor de 0 [24].

Denotado el contorno en la imagen de interés se procedió a crear una función para realizar la segmentación. Para esto se utilizaron una serie de funcionalidades que permitieron extraer el área de interés de cada una de las imágenes de la base de datos. Entre ellas, la función cv2.boundingRect() y la función píxel de OpenCV [25]. La primera función utiliza un rectángulo aproximado alrededor de la imagen binaria. Este rectángulo se utiliza principalmente para resaltar la región de interés después de obtener los contornos de una imagen. La segunda, se utilizó para la creación de una función que permitió recorrer el largo y ancho la célula de interés. Con la función píxel se obtuvo el valor de intensidad de cada uno de los píxeles con base en la posición en *x* y *y.* Utilizando la variable suma, se almacenó el valor de cada píxel obtenido en las iteraciones. Almacenados estos valores, se realizaron comparaciones entre zonas oscuras y zonas con alta saturación permitiendo extraer la célula de interés.
Como paso final en la etapa de pre-procesamiento se extrajeron las características de las células segmentadas.
Tomando en cuenta los resultados obtenidos en el proceso de segmentación se validaron las funciones que permitieron trabajar con las características morfológicas.
Las características utilizadas fueron las siguientes:

- Redondez: Está denominada por las características generales de la forma en lugar de la definición de sus bordes y esquinas o la rugosidad de la superficie de un objeto fabricado [26].

- Excentricidad: Mide la longitud más corta de las trayectorias de un vértice *x* para alcanzar cualquier otro vértice *w* de una gráfica conectada. Para utilizar esta función la imagen debe estar previamente filtrada o en escala de grises [27].

- Solidez: Es la relación entre el área de contorno y el área de casco convexo. Por definición, este valor debe ser menor que 1. Es decir, la cantidad de pixeles dentro de una forma no puede superar en número a la cantidad de pixeles del casco convexo [28].

4.3 Diseño e implementación en Software de un algoritmo de aprendizaje. Con base en el diseño implementado en la etapa de pre-procesamiento se crearon las matrices de características y de etiquetas (Del inglés: *Labels*). A través de la función StandarScaler() se estandarizó el contenido de las matrices de forma tal que los valores atípicos fueran más apropiados para el uso de los clasificadores. Una vez realizada la estandarización de las matrices se aplicó *PCA* con un valor de 3, que hace referencia a los datos con mayor variabilidad, en este caso las características extraídas.
Realizado el proceso anterior se integró con cada uno de los clasificadores. En nuestro caso se utilizó *K-NN*, *SVM lineal* y *Bosque Aleatorio*.

En el *clasificador K-NN* se utilizó la función *clf =KNN (n_neighbors =n_ideal),* la cual permite determinar el mejor valor de n vecinos [29]. Este valor arrojará en términos de eficiencia el mejor resultado de acuerdo con los valores ingresados correspondientes a las matrices de características y de prueba.
Para el *clasificador SVM lineal* se configuró el parámetro C con un arreglo de valores de (0.001,0.01,0.1,1,10) con el finde mitigar el sobreajuste y buscar el mejor desempeño en la clasificación.
En el caso del clasificador ***RF***, se configuró un arreglo con los valores correspondientes a los árboles de decisión cuyos valores fueron (1,5,25,50,75).

4.4 Validación del desempeño de los algoritmos diseñados. Los escenarios propuestos para los resultados de clasificación estuvieron sujetos al uso de reducción de componentes en los clasificadores, así como la implementación de estos algoritmos donde se prescindió de este parámetro.
El primer escenario propuesto se basó en el uso de dos conjuntos de prueba donde se obtuvieron porcentajes correspondientes a exactitud (Del inglés: *Accuracy*), precisión y F1-score. Las muestras utilizadas fueron las siguientes:

- Eosinófilos con leucemia vs Eosinófilos sanos.
- Monocitos con leucemia vs Monocitos sanos.
- Neutrófilos con leucemia vs Neutrófilos sanos.
- Linfocitos con leucemia vs Linfocitos sanos.

El segundo escenario propuesto fue la integración de todos los conjuntos de prueba, es decir, se incluyeron las células con leucemia y las células sanas.

5. Resultados.

Los resultados aquí presentados corresponden al desarrollo de este trabajo en cada una de las etapas mencionadas que se utilizaron en la metodología. En primera instancia, se cargaron las imágenes al algoritmo creado para la clasificación.

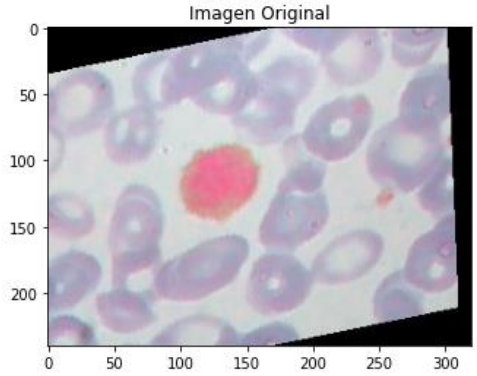


Figura 1. Monocito en paciente enfermo

5.1 Etapa de pre-procesamiento. Se ilustran los resultados obtenidos en cada paso del pre-procesamiento, obteniendo al final, la imagen segmentada.

- Aplicación de ecualización de histograma

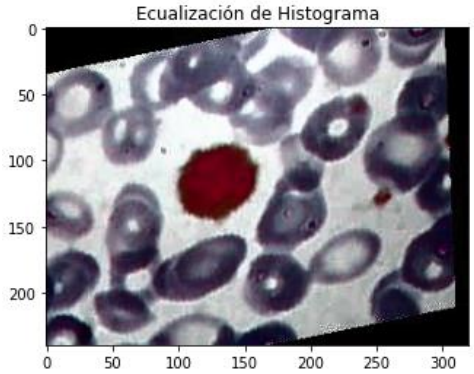


Figura 2. Ecualización de histograma en monocito

- Identificación de canal *S* del espacio *HSV*

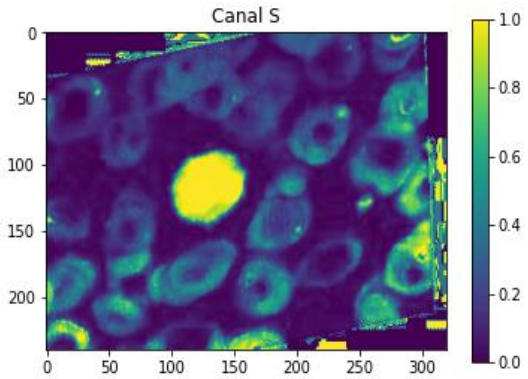


Figura 3. Canal S extraído del espacio HSV

- Aplicación de filtro de media en canal *S*

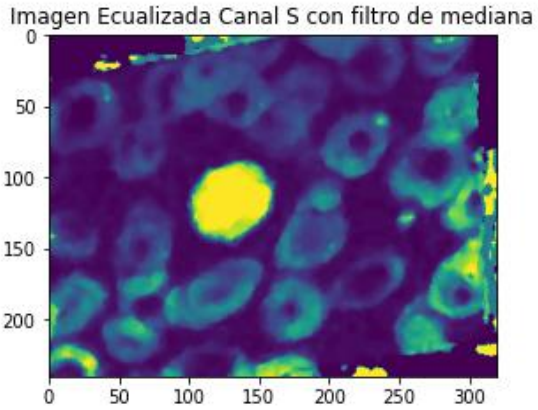


Figura 4. Canal *S* con filtro de madia aplicado

- Detección de bordes (Tresholding)

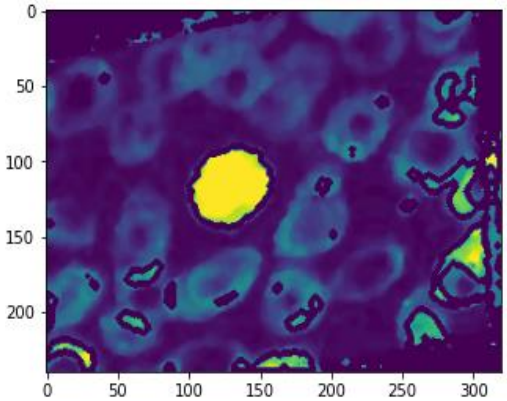


Figura 5. Detección de bordes canal *S* con filtro de media

- Segmentación y extracción de célula de interés.

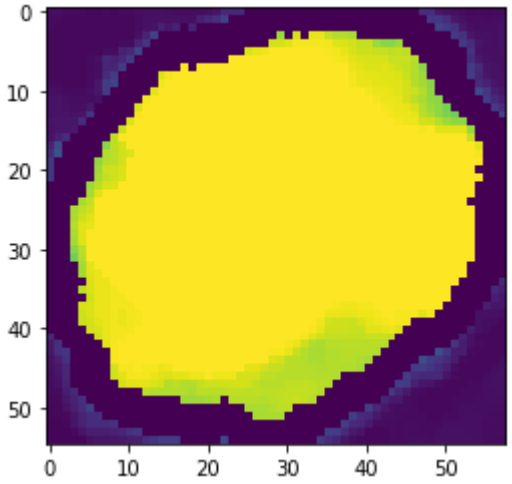
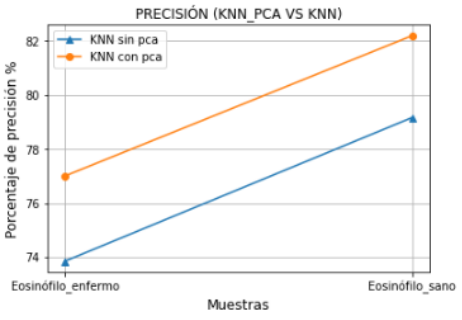


Figura 6. Imagen de célula segmentada

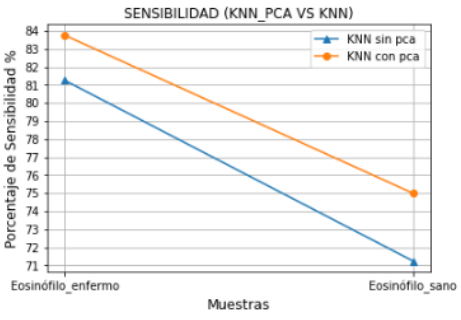
5.2 Etapa de clasificación. Se ilustran las figuras correspondientes a los porcentajes obtenidos en los parámetros de clasificación de cada uno de los algoritmos utilizados en cada uno de los escenarios propuestos.

Los escenarios propuestos se desarrollaron bajo el marco del uso de reducción de dimensiones PCA y el no uso de esta función en cada uno de los clasificadores. Además, se realizó la búsqueda de los hiperparámetros óptimos para cada uno de los clasificadores.

Escenario 1. Comparación Eosinófilo enfermo vs Eosinófilo sano.



A



B

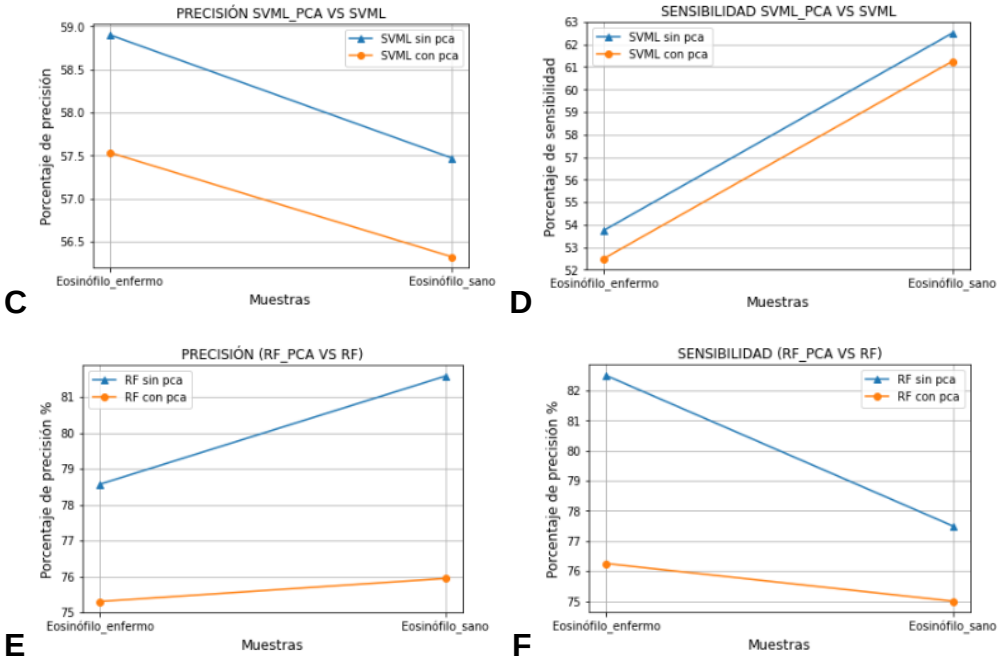


Figura 7. A. Resultado de porcentaje de precisión *K-NN*, B. Porcentaje de sensibilidad *K-NN*, C. Porcentaje de precisión *SVML*, D. Porcentaje de sensibilidad *SVML*, E. Porcentaje de precisión *RF*, F. Porcentaje de sensibilidad *RF* en pacientes con leucemia y sujetos sanos.

Parámetro óptimo *K-NN* con *PCA (Número de vecinos): 3*

Parámetro óptimo *K-NN* sin *PCA* (*Número de vecinos*): *3*

Parámetro óptimo *SVML* con *PCA (Valor de C): 1*

Parámetro óptimo *SVML* sin *PCA (Valor de C): 0,1*

Parámetro óptimo *RF* con *PCA (Valor de árbol): 25*

Parámetro óptimo *RF* sin *PCA (Valor de árbol): 50*

Escenario 2. Monocito enfermo vs Monocito sano.

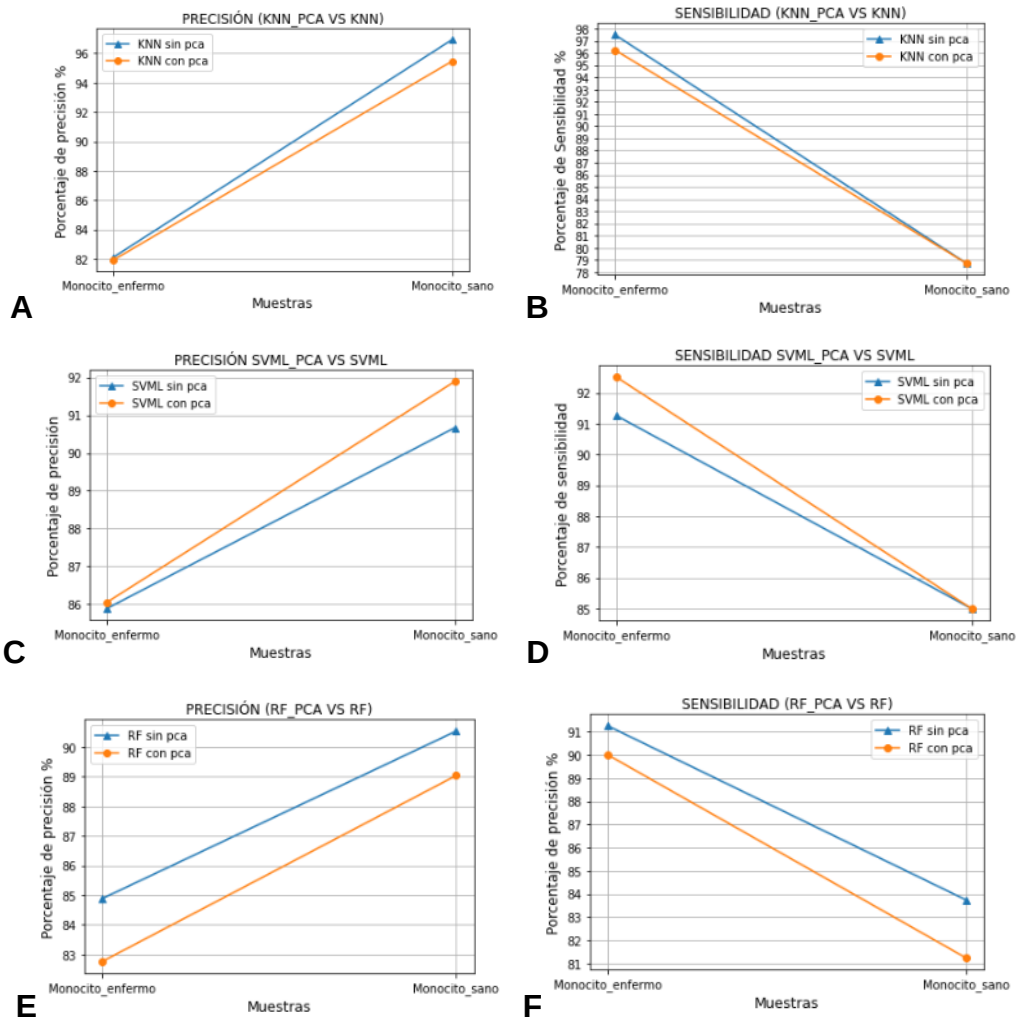


Figura 8. A. Resultado de porcentaje de precisión *K-NN*, B. Porcentaje de sensibilidad *K-NN*, C. Porcentaje de precisión *SVML*, D. Porcentaje de sensibilidad *SVML*, E. Porcentaje de precisión *RF*, F. Porcentaje de sensibilidad *RF* en pacientes con leucemia y sujetos sanos.

Parámetro óptimo *K-NN* con *PCA (Número de vecinos): 3*

Parámetro óptimo *K-NN* sin *PCA* (*Número de vecinos*): *5*

Parámetro óptimo *SVML* con *PCA (Valor de C): 1*

Parámetro óptimo *SVML* sin *PCA (Valor de C): 0,1*

Parámetro óptimo *RF* con *PCA (Valor de árbol): 25*

Parámetro óptimo *RF* sin *PCA (Valor de árbol): 50*

Escenario 3. Neutrófilo enfermo vs Neutrófilo sano

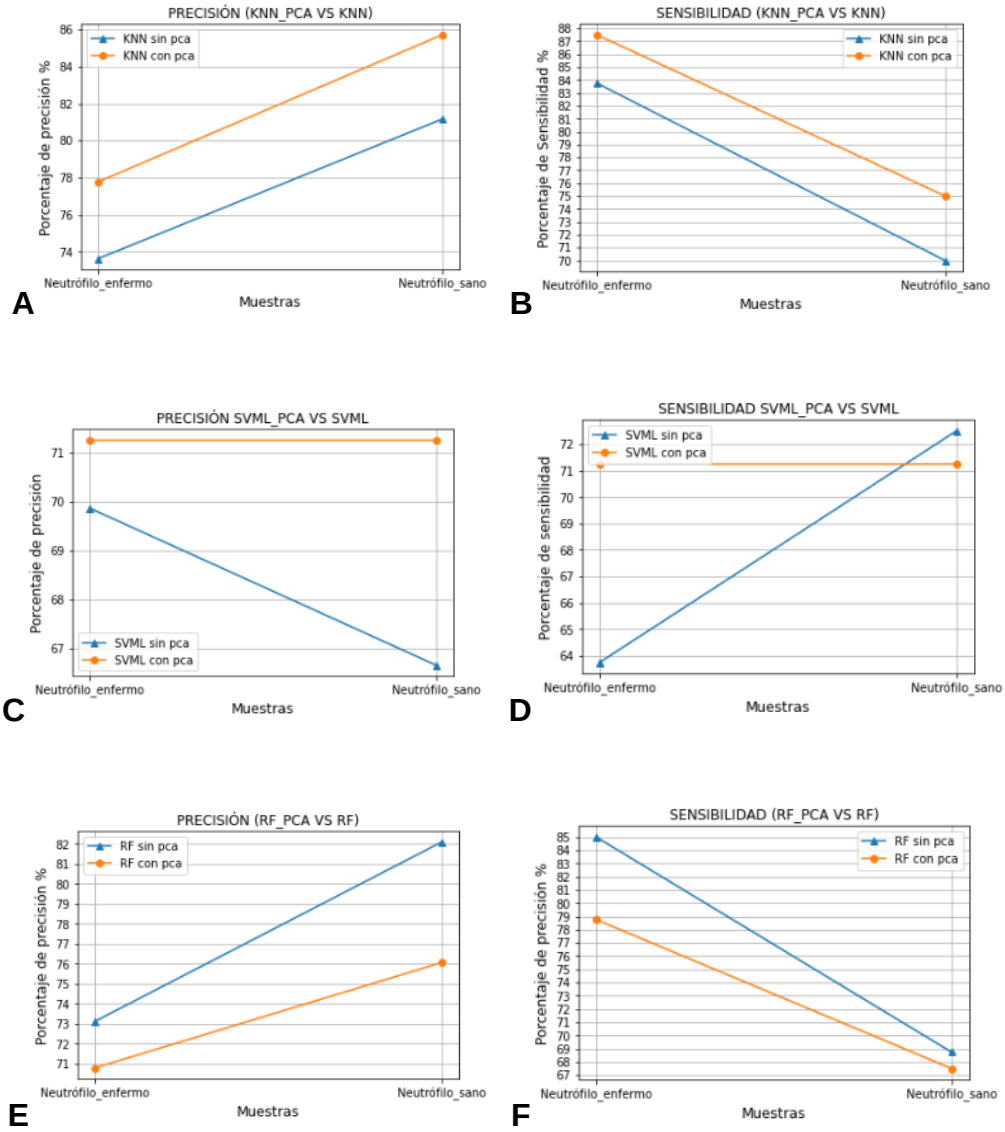


Figura 8. A. Resultado de porcentaje de precisión *K-NN*, B. Porcentaje de sensibilidad *K-NN*, C. Porcentaje de precisión *SVML*, D. Porcentaje de sensibilidad *SVML*, E. Porcentaje de precisión RF, F. Porcentaje de sensibilidad *RF* en pacientes con leucemia y sujetos sanos.

Parámetro óptimo *K-NN* con *PCA (Número de vecinos): 3*

Parámetro óptimo *K-NN* sin *PCA* (*Número de vecinos*): *7*

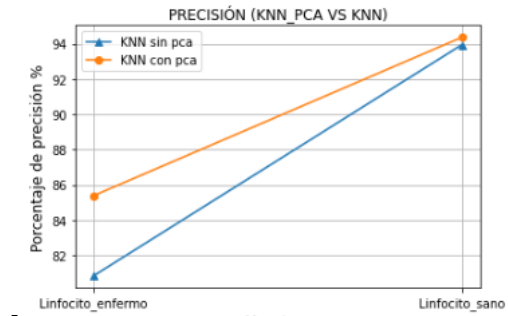
Parámetro óptimo *SVML* con *PCA (Valor de C): 1*

Parámetro óptimo *SVML* sin *PCA (Valor de C): 0,1*

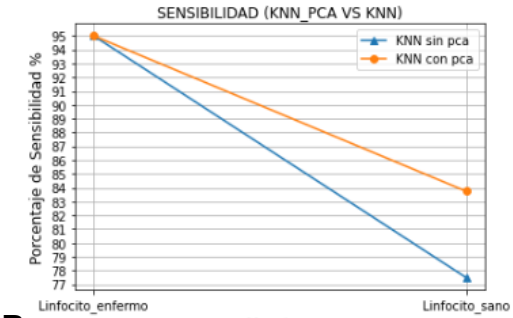
Parámetro óptimo *RF* con *PCA (Valor de árbol): 25*

Parámetro óptimo *RF* sin *PCA (Valor de árbol): 50*

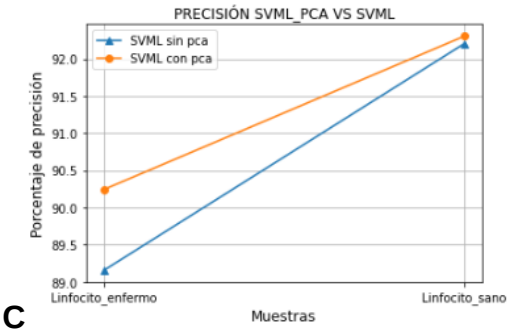
Escenario 4. Linfocito enfermo vs Linfocito sano



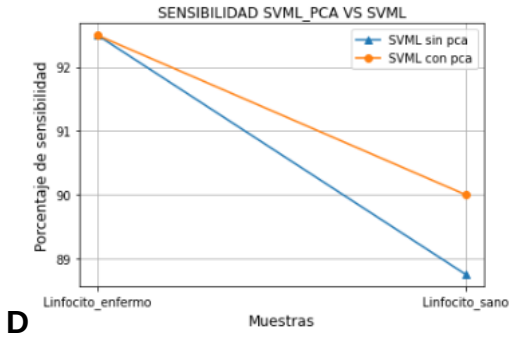
A



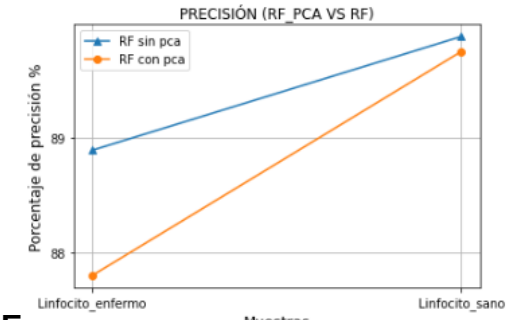
B



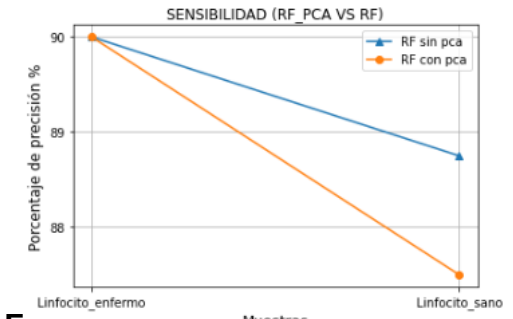
C



D



E



F

Figura 9. A. Resultado de porcentaje de precisión *K-NN*, B. Porcentaje de sensibilidad *K-NN*, C. Porcentaje de precisión SVML, D. Porcentaje de sensibilidad *SVML*, E. Porcentaje de precisión *RF*, F. Porcentaje de sensibilidad *RF* en pacientes con leucemia y sujetos sanos.

Parámetro óptimo *K-NN* con *PCA (Número de vecinos): 3*
Parámetro óptimo *K-NN* sin *PCA* (*Número de vecinos*): *5*
Parámetro óptimo *SVML* con *PCA (Valor de C): 0,1*
Parámetro óptimo *SVML* sin *PCA (Valor de C): 0,1*
Parámetro óptimo *RF* con *PCA (Valor de árbol): 25*
Parámetro óptimo *RF* sin *PCA (Valor de árbol): 75*

Dentro de los resultados obtenidos, la matriz de confusión permite identificar aquellos valores que determinan el desempeño del clasificador utilizado.
Para este trabajó se tomó como escenario, el mejor desempeño obtenido por los clasificadores.

Matriz de confusión. Clasificador *K-NN*

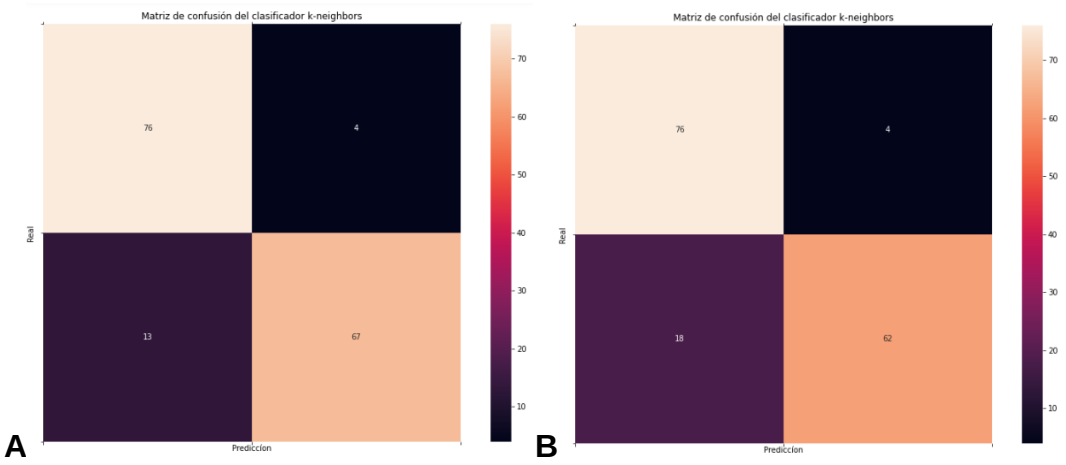


Figura 10. A. Matriz de confusión clasificador *K-NN* con *PCA*, B. Matriz de confusión clasificador *K-NN* sin *PCA.*

Matriz de confusión. Clasificador *SVML*

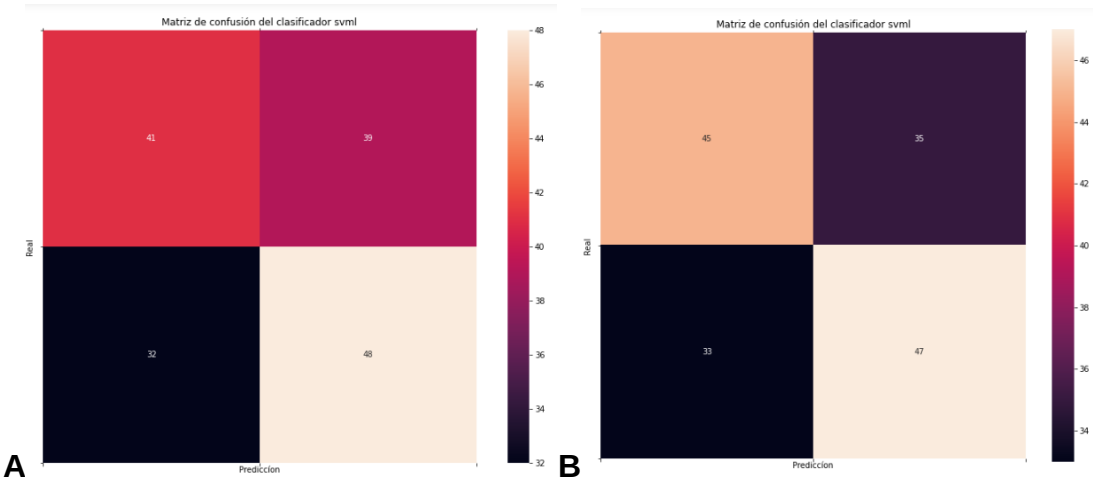


Figura 10. A. Matriz de confusión clasificador *SVML* con *PCA*, B. Matriz de confusión clasificador *SVML* sin *PCA.*

Matriz de confusión. Clasificador *RF*

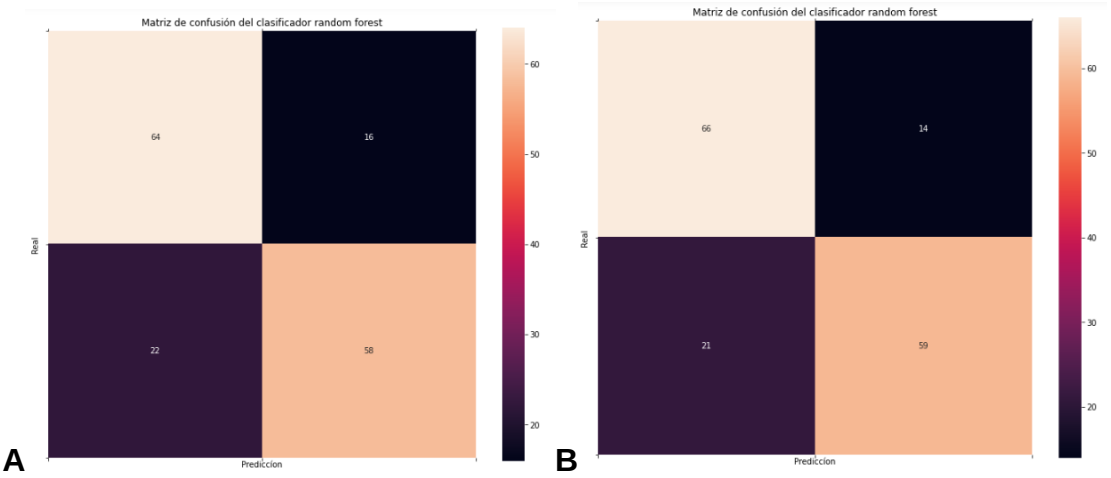


Figura 11. A. Matriz de confusión clasificador *RF* con *PCA*, B. Matriz de confusión clasificador *RF* sin *PCA.*

5.3 Porcentajes de exactitud. Resultado de comparación de porcentajes de exactitud de los algoritmos de clasificación en cada uno de los escenarios propuestos

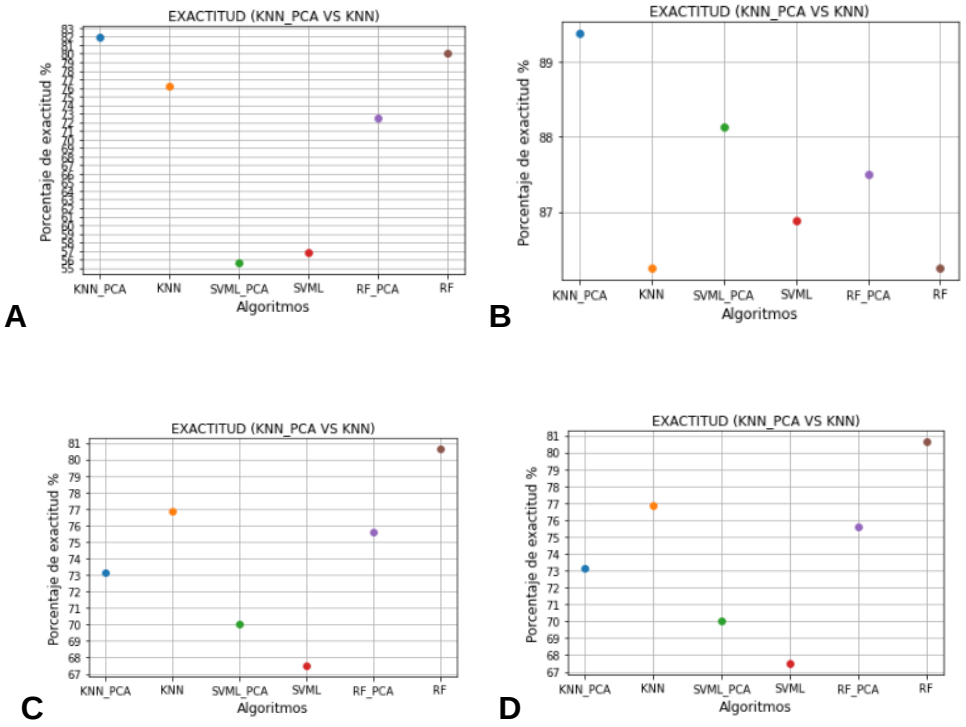


Figura 10. A. Porcentajes de exactitud en escenario 1, B. Porcentaje de exactitud en escenario 2, C. Porcentaje de exactitud en escenario 3, D. Porcentaje de exactitud escenario 4

Se puede observar como el algoritmo *K-NN* obtiene mejor porcentaje de exactitud en la imagen A y la imagen B. A pesar de que decae su exactitud en los siguientes escenarios C y D, permanece en un umbral alto, comparado con *SVML* donde no se llega a porcentajes superiores al 80%. El clasificador *RF* presenta buenos resultados de exactitud, sin embargo, su desempeño requiere de altos de cómputo recursos para obtener un buen desempeño.

5.4 Discusión de resultados. En primer lugar, el proceso de segmentación se desarrolló tomando como base la literatura encontrada. Sin embargo, durante la elaboración del código, se eliminó el uso de imágenes en escala de grises. La principal razón es que se tuvo claro el uso del canal *S* extraído del espacio de colores HSV y la representación en gris no generó cambios significativos que indicaran su uso.

Con respecto a la función que se creó para la segmentación, se realizó un proceso de extracción de las células previo al proceso de clasificación. Con estas pruebas se quiso ver la eficiencia segmentación. En general el funcionamiento fue bueno, ya que se logró extraer las células de interés de la gran mayoría de las imágenes de la base de datos y sólo en algunas, se obtuvo valores altos de saturación que iban más allá de los bordes de la célula de interés y por ende extrayendo regiones no deseadas.

En general, el algoritmo de *K-NN* sin uso de PCA obtuvo mejores valores de precisión. El uso de *PCA*, es eficiente cuando se tiene una gran cantidad de muestras de estudio, lo cual no aplica para este caso ya que solo se utilizaron cuatro tipos de células y de cada una se extrajeron tres características morfológicas, lo que termina no siendo representativo al utilizar reducción de dimensiones. Sin embargo, los resultados altos de precisión se dieron con las células sanas. A nivel clínico, los tintes utilizados, resaltan el grupo de interés, denotando la forma característica de las células. Debido a que las células cancerígenas presentan anomalías a nivel morfológico, en términos de procesamiento de imágenes aportan ruido, información que inclusive luego de filtrada permanecerá como parte de los resultados obtenidos en los procesos de segmentación y que se reflejará en los algoritmos de clasificación.
A razón de lo anterior, sólo para el caso de los monocitos hubo un alto porcentaje de acierto en los tres clasificadores. Una causa de esto es que los monocitos debido a su bajo contenido granular poseen una superficie más lisa que el resto de las células.

A diferencia de los resultados obtenidos con los eosinófilos, los tres clasificadores obtuvieron altos porcentajes de precisión y sensibilidad identificando monocitos. En este caso, ambos clasificadores *K-NN* y *RF* presentaron resultados similares identificando monocito sano sin utilizar *PCA.* Otro aspecto importante es el alto reconocimiento de células sanas por parte del clasificador *SVML.* En el primer escenario obtuvo un porcentaje de precisión no superior al 60%, sin embargo, para el segundo escenario este porcentaje aumentó obteniendo mejoras cuando se utilizó con *PCA.* Una razón para este comportamiento puede ser que, al utilizar reducción de dimensiones, el clasificador haya encontrado clases linealmente separables las cuales dependen de las características extraídas que se utilizaron en el escenario planteado. Como método para mejorar los

porcentajes de precisión, se plantea el uso un clasificador basado en hiperplano que, aunque no separa las clases de manera perfecta, sea más robusto y tenga mayor capacidad de predicción al aplicarlo en nuevas observaciones. Una estrategia consiste en realizar variaciones de *C,* la cual determina la penalización por violar el margen impuesto por el hiperplano.

A diferencia del escenario anterior, el porcentaje de precisión más alto se dio utilizando el clasificador *K-NN.* Una contribución a este resultado es que, entre las características encontradas, la excentricidad da información de la cantidad de esquinas y superficie rugosa de la célula de estudio.

En cuanto al clasificador *RF,* aunque presenta buen desempeño, su tiempo de procesamiento es más lento comparado con *K-NN.* En este caso, puede ocurrir, Sin embargo, se mejora su rendimiento si las características extraídas se ajustan aplicando otro tipo de preprocesamiento, como por ejemplo el uso de transformada *RGB,* y posteriormente.

Se observa que el porcentaje de precisión fue más alto utilizando los clasificadores *K-NN* y *RF.* Sin embargo, ambos difieren en cuanto a que el primer clasificador obtuvo mejor porcentaje cuando se aplicó *PCA*. Esto se debe a que *K-NN* clasifica valores buscando los puntos de datos más similares por cercanía que aprendió en la etapa de entrenamiento requiriendo para esto menos recursos de memoria y bases de datos pequeños sin una cantidad grande de características.

Se observa un alto porcentaje de exactitud entre los clasificadores *K-NN* y *RF.* Sin embargo, ninguno de estos valores supera el 90%. Lo anterior indica que es necesario, tener en cuenta los porcentajes de precisión para estimar el clasificador que arroja mejores resultados de clasificación.

6. Conclusiones.

Este trabajo se enfocó en diseñar y aplicar un algoritmo de clasificación que permitiera la detección de blastos de leucemia. Dentro de este proceso se incluyó el análisis de imágenes, la implementación de operaciones morfológicas y el uso de técnicas de pre-procesamiento y clasificación de imágenes.

Durante el desarrollo de este proyecto se aplicaron diversas técnicas de pre-procesamiento como la conversión de espacio de colores para presentar valores más saturados en la célula de interés; aplicando luego filtro de media para eliminar las componentes no deseadas preservando los bordes. Se segmentaron las células de interés y se procedió con la extracción de características morfológicas. Se evaluaron tres clasificadores de acuerdo con lo establecido en la etapa de pre-procesamiento y se procedió con la validación de resultados.

Se propusieron escenarios realizando comparación entre células de la misma clase diferenciando entre enfermas y sanas. El más complejo de clasificar se obtuvo por parte de *SVML* ya que el parámetro C es más sensible a cambios realizados en su inicialización que en consecuencia fueron determinantes ya que su funcionamiento con respecto a *K-NN* y *RF* fue peor al momento de evaluar la exactitud y la precisión.

Los algoritmos con mejor desempeño fueron *K-NN* y *RF.* Ambos demostraron buenos porcentajes de precisión y exactitud, sin embargo, se elige el clasificador *K-NN* ya los porcentajes de precisión obtenidos estuvieron por encima del 80%, además de adaptarse rápido número de muestras y características extraídas con menos recursos de cómputo, y su inicialización en la cantidad de vecinos para su desempeño requirió de un grupo máximo de 7 vecinos.

Con base en los escenarios propuestos y los resultados obtenidos se concluye que la tarea de extracción de glóbulos blancos en un extendido de sangre periférica sigue siendo un reto para los algoritmos de clasificación, esto debido a las características morfológicas con las que cuenta cada grupo de estudio, los detalles lumínicos, técnicas de pre-procesamiento utilizadas y la evaluación del clasificador utilizado.

6.1 Trabajo futuro. Como propuesta para el mejoramiento del algoritmo de segmentación se proponen varias tareas que pueden ayudar a un mejor diagnóstico:

- Proponer un algoritmo de segmentación que permita separar la membrana de las células del núcleo.

- Definir un nuevo grupo de características de tipo morfológicas que permitan utilizar subsegmentos que conforman a las células.

- Obtener un mayor número de imágenes por paciente que permitan probar distintos clasificadores y con base en los porcentajes de acierto tomar decisiones más acertadas en cuanto al que mejor se adopta a las tareas de clasificación.

- Construir una amplia base de datos con células sanas por medio de sangre periférica, ya que por temas de costo es más viable que recolectar la información por cada paciente.

7. Referencias

[1]. Jaffe, E. S. (Ed.). (2001). Pathology and genetics of tumors of haematopoietic and lymphoid tissues (Vol. 3). Iarc.

[2]. Tonato Lovato, P. C. (2017). *Leucemia linfoide* (Bachelor's thesis, Universidad Técnica de Ambato-Facultad de Ciencias de la Salud-Carrera de Medicina).

[3]. Neri Guarachi, L. E., Torres Aldunate, T. G., & Gina, T. *Caracterización epidemiológica, demográfic hematológica de las leucemias mieloides agudas por citometría de flujo en las gestiones 2011-2012* (Doctoral dissertation).

[4]. S. Khalid, T. Khalil, and S. Nasreen, “A survey of feature selection and feature extraction techniques in machine learning,” in Science and Information Conference (SAI), 2014, 2014, pp. 372– 378.

[5]. Guadarrama, K. Z. H. (2007). *SEGMENTACIÓN DE CÉLULAS DE LEUCEMIA UTILIZANDO TÉCNICAS DE CLASIFICACIÓN* (Doctoral dissertation, BENEMÉRITA UNIVERSIDAD AUTÓNOMA DE PUEBLA).

[6]. Peña Serrato, O. M. Análisis de la tipificación de leucemias por citometría de flujo en la Pontificia Universidad Javeriana.

[7]. https://amoxkali.cs.buap.mx/archivo/TES223.pdf

[8]. (Pino, D., Macías Abrahm, C., Lahera Sánchez, T., Marsán Suárez, V., Sánchez Segura, M., del Valle, L., Socarras Ferrer, B., & Martínez Machado, M. (2013). Caracterización inmunofenotípica de pacientes con leucemia mieloide aguda. *Revista Cubana de Hematología, Inmunología y Hemoterapia, 30*(1).

[9]. Zapata-Tarres M, Sánchez-Huerta JL, Angeles-Floriano T, et al. Identificación de alteraciones moleculares en pacientes pediátricos con diagnóstico de leucemia aguda. Rev Hematol Mex. 2017;18(2):47-57.

[10]. http://www.cancerinfo.es/index.php?textoid=21&orden=4

[11]. http://inaoe.repositorioinstitucional.mx/jspui/handle/1009/435

[12]. https://pythondiarioa.com/2018/01/ontroduccop-al-machine-learning-7-Ios.html

[13]. Dharani T ME/CS Department of ECE Saranathan College of Engineering Trichy, India, Hariprasath S Assistant Professor Department of ECE Saranathan College of Engineering Trichy, India.: Diagnosis of Leukemia and its types Using Digital Image Processing Techniques. IEEE Trans. Inf. Theory 275-279.

[14]. J. Laosai and K. Chamnongthai, “Acute leukemia classification by using SVM and K-Means clustering,” in Proceedings of the 2014 International Electrical Engineering Congress, iEECON 2014, Thailand, March 2014.

[15]. G Biau. 2012. Analysis of a random forests model. J. Mach. Learn. Res. 13, (2012), 1063– 1095.

[16]. Kan Jiang; Qing-Min Liao; Sheng-Yang Dai, “Machine Learning and Cybernetics”, 2003 International Conference on Volume 5, Issue, 2-5 Nov. 2003 Page(s): 2820 - 2825 Vol.5.

[17]. Carlos Platero, Análisis de imágenes biomédicas en color procedente de microscopía en campo claro, Grupo de Bioingeniería Aplicada, Dpto. Electrónica, Automática e Informática Industrial Universidad Politécnica de Madrid, JOURNAL OF VIROLOGY

[18]. W. Srisukkham, L. Zhang, S. C. Neoh, S. Todryk, and C. P. Lim, “Intelligent leukaemia diagnosis with bare-bones PSO based feature optimization,” Appl. Soft Comput., vol. 56, pp. 405–419, 2017.

[19]. Dharani, T., & Hariprasath, S. (2018, October). Diagnosis of leukemia and its types using digital image processing techniques. In *2018 3rd International Conference on Communication and Electronics Systems (ICCES)* (pp. 275-279). IEEE.

[20]. Ahmed M. Abdeldaim, Ahmed T. Sahlol, Mohamed Elhoseny and Aboul Ella Hassanien.: Computer-Aided Acute Lymphoblastic Leukemia Diagnosis System Based on Image Analysis. Culture & Science City, 6th of October 15525, Egypt, pp. 131-147 (2018).

[21].http://acodigo.blogspot.com/2017/08/histogramas-opencv python.html#:~:text=La%20ecualizaci%C3%B3n%20de%20histograma%20busca,nos%20provee%20la%20funci%C3%B3n%20cv2.

[22].https://medium.com/@gastonace1/detecci%C3%B3n-de-objetos-por-colores-en-im%C3%A1genes-con-python-y-opencv-c8d9b6768ff.

[23]. [17]. https://ichi.pro/es/filtros-de-imagen-en-python-42806119651282.

[24]. https://docs.opencv.org/4.x/d7/d4d/tutorial_py_thresholding.html.

[25]. https://www.pythonpool.com/cv2-boundingrect/.

[26]. https://es.acervolima.com/mahotas-redondez-de-imagen/.

[27]. https://es.acervolima.com/mahotas-excentricidad-de-la-imagen/.

[28]. https://unipython.com/propiedades-los-contornos/.
[29].https://www.aprendemachinelearning.com/clasificar-con-k-nearest-neighbor-ejemplo-en-python/.
[30]. https://programmerclick.com/article/8860223797/.