Obtención de Características de Subtipos de Leucemia en Imágenes Digitales de Células Sanguíneas para su Clasificación

por

Martha Coral Galindo Domínguez

Tesis sometida como requisito parcial
para obtener el grado de

MAESTRA EN CIENCIAS
EN LA ESPECIALIDAD DE
CIENCIAS COMPUTACIONALES

en el
Instituto Nacional de Astrofísica, Óptica y Electrónica.

Supervisada por:

DR. JESÚS ANTONIO GONZÁLEZ BERNAL
Coordinación de Ciencias Computacionales, INAOE

Tonantzintla, Pue.
2008

Abreviaturas

AcMo	Anticuerpos monoclonales
ADN	Ácido desoxirribonucleico
B	Banda Azul
CIE	Comisión Internacional de l'Eclairage
CMF	Citometría de flujo
EUA	Estados Unidos de América
FAB	Grupo Cooperativo Franco-Americano-Británico
G	Banda Verde
HSV	Modelo de color
IMSS	Instituto Mexicano del Seguro Social
L_1	Subtipo 1 de Leucemia Aguda Linfoblástica
L_2	Subtipo 2 de Leucemia Aguda Linfoblástica
LAL	Leucemia Aguda Linfoblástica
LAM	Leucemia Aguda Mieloblástica
M_2	Subtipo 2 de Leucemia Aguda Mieloblástica
M_3	Subtipo 3 de Leucemia Aguda Mieloblástica
M_5	Subtipo 5 de Leucemia Aguda Mieloblástica
OMS	Organización Mundial de la Salud
ACP	Análisis de Componentes Principales
R	Banda Roja
ROI	Regiones de Interés
RSC	Recuento sanguíneo completo

TC o TAC	Tomografía computarizada
VC	Visión por computadora

Lista de Figuras

Lista de Tablas

Resumen

A pesar de los recientes adelantos en técnicas hematológicas, incluso en estudios de Citometría de flujo, Inmunofenotipo y ADN, el análisis morfológico del frotis de la sangre periférica o médula ósea continúa siendo una importante investigación inicial en pacientes que sufren trastornos hematológicos.

El reconocimiento de los subtipos de Leucemia aguda en células sanguíneas es importante debido a su uso en el diagnóstico clínico. La identificación de estos subtipos de leucemia aguda en imágenes digitales de células sanguíneas ayudará al médico en la prescripción del tratamiento adecuado del paciente.

Este trabajo presenta un método de generación de características descriptivas para la identificación y clasificación de células sanguíneas de subtipos de leucemia aguda en imágenes digitales. La primera parte consiste en un pre-procesamiento para segmentar la imagen por color para posteriormente detectar los bordes de las células en las tres bandas de la imagen: R, G y B, y de esta manera obtener características de textura, geométricas, estadísticas y la valores propios (ACP) con un 80% de variabilidad. Las características obtenidas se utilizaron como atributos de entrada para realizar el proceso de minería de datos (usando diferentes clasificadores) para reconocer cinco diferentes tipos de células sanguíneas. Como la cantidad de ejemplos de cada subtipo de leucemia aguda de la base de datos está muy desbalanceada, se utilizaron técnicas de sobre-muestreo

[65] para reducir este problema. La evaluación de los resultados se hizo con los expertos del dominio Dr. José E. Alonso Chávez, Dr. Rubén Lobato Tolama y la química Laura O. Olvera Oropeza del Instituto Mexicano del Seguro Social (IMSS) San José en Puebla, y utilizando la técnica de validación cruzada con significancia estadística. Los resultados para cada subtipo de leucemia aguda son superiores al 85% de precisión.

Como resultado se obtuvo un conjunto de características que describen a los subtipos de leucemia aguda y nos permitió clasificarlos con una precisión global de 88%. Logrando una exactitud del 85% para los subtipos L1 y L2 y 91% para los subtipos M2, M3, y M5. Superando el promedio de clasificación realizada por el especialista, cuyo error está entre el 20 y 30% [40].

Abstract

In spite of the recent advances in hematological techniques such as flux cytometry (with immunophenotype), and DNA analysis; morphological analysis of bone marrow smears (even of peripheral blood) are still the starting point to detect patients that suffer of blood disorders. This is why the identification of acute Leukemia subtypes from blood cells is an important task due to its use in clinical diagnosis. The classification of these leukemia subtypes from digital images of blood cells helps the physician to prescribe a suitable treatment to the patient.

This work presents a method to generate descriptive characteristics for the identification and classification of acute Leukemia subtypes from digital images of blood cells. The first part of this work consists of a pre-processing phase to segment the image by color to then detect the boundaries of the cells using the three bands of the image: R, G, and B. In the second phase we use the preprocessed images to obtain their descriptive characteristics: texture, geometric, statistical, and their eigenvalues (ACPs) with 80% of variability. These characteristics were used as input attributes to perform the data mining process (using different classifiers) to recognize five different leukemia subtypes. Since our leukemia database presented the class imbalance problem (because of the different proportion of cases of each leukemia subtype), we applied over-sampling techniques to reduce its impact. The evaluation of the results was done by the domain experts Dr. José E. Alonso Chávez, Dr. Rubén Lobato Tolama, and the chemistry Laura O.

Olvera Oropeza from the “Instituto Mexicano del Seguro Social” (IMSS) San Jose in Puebla. We also performed a quantitative evaluation using the cross validation technique. Our results for each leukemia subtype were around 85% of accuracy.

As result we obtained a set of descriptive characteristics to describe acute leukemia subtypes that allowed us to classify them with a global precision of 88 %. We achieved an accuracy of 85% for subtypes L_1 and L_2 and 91% for subtypes M_2, M_3, and M_5. With these results we outperformed the average classification accuracy obtained by domain experts, whose error ranges from 20% to 30% [40].

Agradecimientos

A Dios, la Virgen y San Judas Tadeo por nunca abandonarme, proporcionarme buena salud y concederme terminar un proyecto más en mi vida.

A mi esposo Héctor y mis hermosas hijas Andy y Dafy por todo su amor, su confianza, y por ayudarme a concluir esta nueva meta, sin ustedes no lo hubiera logrado.

A mis padres Julio y Nila y mis hermanas Lau, Ale, Soni y Karen por su apoyo y ayuda incondicional.

A mi asesor Dr. Jesús A. González por depositar su confianza en mí, por haberme guiado y apoyado en el transcurso de la tesis, pero sobre todo el conocimiento que ha compartido conmigo durante el desarrollo de este trabajo.

A los doctores Manuel Montes y Gómez, Eduardo F. Morales Manzanares y Carlos A. Reyes García, por el tiempo dedicado a la revisión de esta tesis, así también como sus valiosos comentarios.

Al Dr. Iván Olmos, Dr. José E. Alonso Chávez, Dr. Rubén Lobato Tolama y la química Laura O. Olvera Oropeza del Instituto Mexicano del Seguro Social

(IMSS) San José en Puebla por su ayuda en conocimientos médicos y por el material facilitado para el desarrollo de este trabajo.

A mis amigos Chayito, Rosy, Paty, Nadia, Alberto, Gustavo, Erika y Javier por sus consejos y su apoyo académico.

Al Consejo Nacional de Ciencia y Tecnología CONACYT) por el apoyo otorgado con la beca No. 202001 y al Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE) por la formación académica.

Dedicatorias

Al amor de mi vida mi esposo Héctor

A mis hijas Andy y Dafy

A mis padres

Tabla de Contenido

Capítulo 1

Introducción

El cáncer es un problema de salud pública a escala mundial, así lo demuestran sus altas tasas de incidencia y mortalidad. En Estados Unidos de América (EUA), el cáncer es la segunda causa de muerte después de los ataques cardiacos, en Latinoamérica el cáncer ocupa el tercer lugar de las causas de muerte [20, 30, 47].

En México de acuerdo con las estadísticas en día Mundial contra el Cáncer, realizadas por el INEGI [12] en el año 2005, existe un alto porcentaje de ingresos hospitalarios por cáncer en la sangre (leucemia) con 49.7% para hombres y para mujeres el 18.4%, así como una taza de mortalidad de 6.3% para hombres y 5.5% para mujeres como se muestra en la figura 1.1. En la población de 1 a 4 años de edad, representó 52 de cada 100 defunciones, y en el grupo de 5 a 14 años, provocó el 55.8% de los fallecimientos.

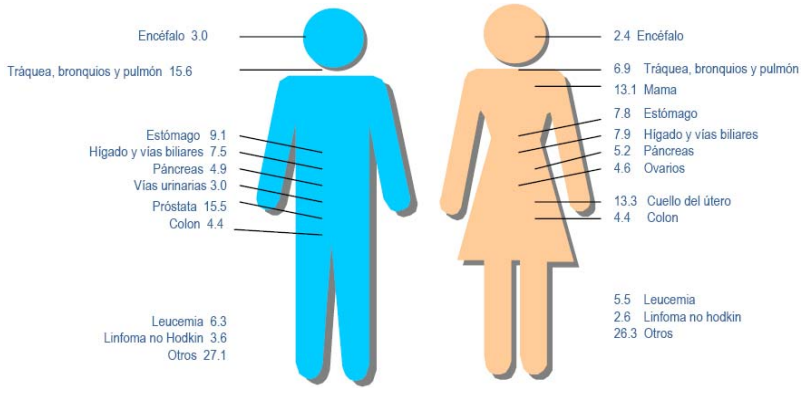


Fig. 1.1 Distribución porcentual de las principales causas de defunción por tumores malignos según sexo, 2005

La leucemia aguda linfoide (LAL) es más frecuente en niños, constituye aproximadamente el 25 por ciento de todos los cánceres en niños menores de 15 años. La leucemia mieloide afecta más a los adultos, con una incidencia creciente.

La detección oportuna, diagnóstico adecuado y tratamiento eficiente de cáncer se encuentran entre los principales problemas de salud en el país.

Aunque no se conoce alguna forma de prevenir la leucemia, si se puede tratar eficazmente con tratamientos modernos. Para el médico es muy importante diagnosticar la leucemia en sus primeras etapas ya que el tratamiento tiene mayor probabilidad de ser exitoso.

Para detectar la leucemia existen diversos procedimientos o análisis como son: examen físico y antecedentes, recuento sanguíneo completo (RSC) [38], frotis de sangre periférica, aspiración de la médula ósea y biopsia, análisis citogenético, inmunofenotipo y tomografía computarizada (TC o TAC) [43]. Algunas de estas técnicas pueden ser complementarias. Desafortunadamente muchas de las técnicas recientes sólo están disponibles en muy pocos laboratorios y el costo de estos análisis es muy alto.

Debido a lo anterior el examen morfológico en el frotis de la sangre periférica o médula ósea continúa siendo una investigación inicial importante en pacientes que sufren trastornos hematológicos, por su bajo costo y porque se puede realizar en muchos laboratorios.

La leucemia es una enfermedad de la sangre que produce un aumento incontrolado de glóbulos blancos (cáncer hematológico) y puede provocar infecciones, anemia y sangrado excesivo. Las células de leucemia viven mucho más, actúan y son diferentes a las células sanguíneas normales y no

pueden realizar la función prevista (defender al organismo de agentes patógenos) [9]. Cuando los glóbulos blancos no llegan al nivel de madurez requerido por el organismo, éste no es capaz de protegerse entonces se dice que hay leucemia [8].

Hay diferentes tipos de leucemias y estas se clasifican por como progresa la enfermedad y por la célula que afecta. La leucemia se presenta en formas agudas y crónicas refiriéndose a como se desarrolla y progresa la enfermedad, donde las leucemias agudas evolucionan rápidamente mientras que las crónicas de manera gradual. Con respecto a la célula que afecta existen dos familias principales de leucemia, la mieloide y linfoide, reciben estos nombres por el tipo de glóbulo blanco afectado, ya sea neutrófilo ó linfocito respectivamente.

La leucemias agudas se clasifican morfológicamente de acuerdo al Grupo Cooperativo Franco-Americano-Británico (FAB) [5] en subtipos y cada uno de ellos presenta distintas características de la enfermedad; por lo tanto, hay diferentas opciones de tratamiento dependiendo del subtipo de leucemia.

El presente trabajo se enfoca en las leucemias agudas mieloblásticas y linfoblásticas.

1.1 Motivación

El análisis morfológico de la sangre consiste primero en contabilizar los glóbulos (blancos y rojos), las plaquetas, y medir la hemoglobina de la sangre. Después se examinan morfológicamente las diferentes células y se determina si son anormales, este procedimiento es realizado por un químico o un hematólogo con experiencia en la detección de la enfermedad (el

análisis morfológico es difícil y de ahí que requiera de personas experimentadas para realizarlo). 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 y de su estado de cansancio.

Por otro lado, un análisis morfológico automatizado para detectar subtipos de leucemia sólo requiere las imágenes digitales tomadas del frotis de sangre de médula ósea, las cuales serán analizadas objetivamente ya que no están relacionadas con el cansancio o experiencia del especialista y se podrían reducir los costos así también como mejorar la exactitud; además que puede ser usado remotamente en lugares de bajos recursos. También puede utilizarse para entrenar estudiantes de hematología o de otras áreas afines.

La dificultad de la tarea hace que un método automatizado para el reconocimiento y clasificación de subtipos de leucemia aguda en imágenes digitales de células sanguíneas de medula ósea sea necesario, para la ayuda en el diagnóstico y prescripción del tratamiento. Aquí es donde enfocamos nuestra investigación, en el descubrimiento temprano de subtipos de leucemia aguda a partir de características morfológicas y estadísticas. Aunque otros métodos de descubrimiento de leucemia existan, el estudio morfológico continúa siendo importante porque se pueden hacer con bajo costo y sin equipo especializado.

La clasificación de tipos de leucemias, por medio de la morfología de la célula en imágenes digitales, ya es un problema medianamente resulto, pero es indispensable la clasificación de subtipos de leucemia aguda, porque con un tratamiento correcto pueden existir posibilidades de supervivencia prolongada o curación, ya que este tipo de leucemias se desarrollan rápidamente.

1.2 Objetivo General

El objetivo de este trabajo es la clasificación automática de subtipos de leucemia aguda con una mayor precisión que la manual. Este proceso computarizado le va a ofrecer una herramienta al médico que utilizará como apoyo al diagnóstico médico y le permitirá elegir oportunamente el tratamiento adecuado para el paciente, mejorando la confiabilidad del análisis, el diagnóstico y agilizando la forma de diagnosticar del químico o hematólogo. Así también la combinación de diferentes tipos de características, para la obtención de una mayor precisión en la clasificación.

1.2.1 Objetivos Particulares

- Seleccionar un conjunto de imágenes digitales de células sanguíneas con leucemia para obtener los conjuntos de entrenamiento y de prueba.
- Obtener las características que describan los subtipos de leucemia para cada familia.
- Seleccionar un ensamble para utilizar como clasificador.
- Entrenar el clasificador (ensamble) con el conjunto de entrenamiento.
- Validar la precisión del clasificador con el conjunto de prueba.
- Validar los resultados con el experto en el dominio

1.3 Metodología

Para resolver este problema médico se requiere de un clasificador que pueda distinguir entre los subtipos de leucemia (en cada familia de leucemia);

utilizando técnicas de visión por computadora para segmentar las células y obtener regiones de interés y aprendizaje automático supervisado para su clasificación automática. En el proceso de segmentación se tiene el reto de poder separar las células sanguíneas con leucemia de los distintos frotis aún cuando la tinción y calidad de los mismos sean diferentes. En el proceso de clasificación requerimos de un algoritmo que sea capaz de distinguir entre los cinco subtipos de leucemia y esto se dificulta porque algunos de ellos comparten muchas características. Para lograr una clasificación correcta, se necesita obtener características de las células sanguíneas (glóbulos blancos) que permitan discernir entre los cinco subtipos de leucemia. Para esto, se precisa de algoritmos de segmentación para obtener Regiones de Interés (ROI) y posteriormente extraer características de ellas para alimentar al clasificador (o ensamble de clasificadores), como se muestra en la Figura 1.2.

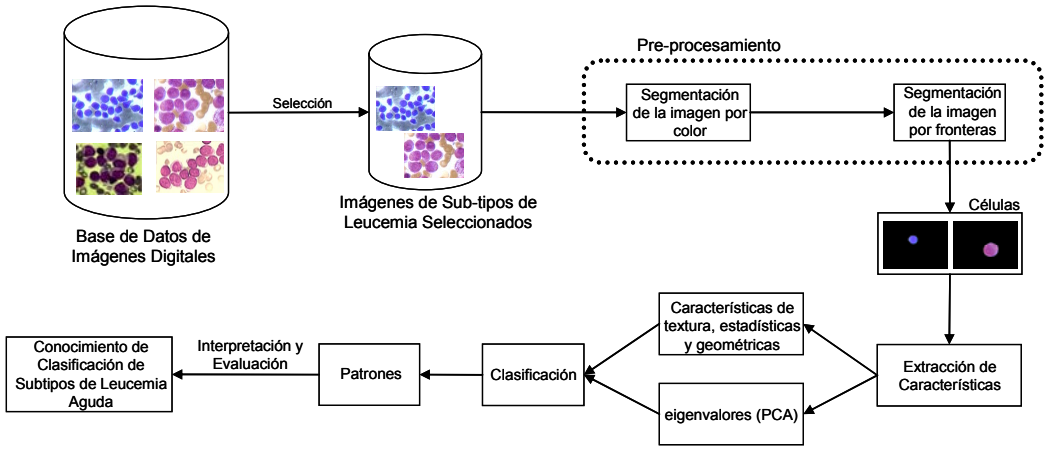


Fig. 1.2 Metodología Propuesta de Clasificación de Subtipos de Leucemia Aguda

1.4 Organización de la Tesis

El presente documento se divide en los siguientes capítulos:

El capítulo 2 describe la estructura de la célula, los distintos tipos de células sanguíneas, así también como los diferentes análisis para detectar leucemia. En el capítulo 3 se explican los fundamentos del análisis y procesamiento de imágenes digitales. En el capítulo 4 se describe el método propuesto para realizar la segmentación y clasificación de regiones de interés. En el capítulo 5 se presentan los resultados obtenidos. Finalmente en el capítulo 6 se mencionan las conclusiones del trabajo realizado y se expone el trabajo futuro.

Capítulo 2

Marco Teórico

En este capítulo se muestra la estructura de la célula y los diferentes tipos de células que existen en la sangre para poder entender mejor el problema explorado. Así mismo, se mencionan los distintos tipos de análisis realizados para poder detectar leucemia. La organización de este capítulo es la siguiente. En la sección 2.1 muestra la anatomía de la célula, después la sección 2.2 presenta los componentes de la sangre, en la sección 2.3 se describe la clasificación de las leucemias de acuerdo a la FAB [5], y por último la sección 2.4 contiene una descripción del diagnóstico de las leucemias

2.1 Anatomía de la célula

Todos los organismos vivos están formados por células; la célula es la unidad mínima que tiene todo ser vivo, y actúa de manera autónoma, a partir de ésta los organismos pueden realizar sus funciones vitales.

La célula esta formada por: membrana celular, el citoplasma y el núcleo, como se muestra en la Figura 2.1 [27].

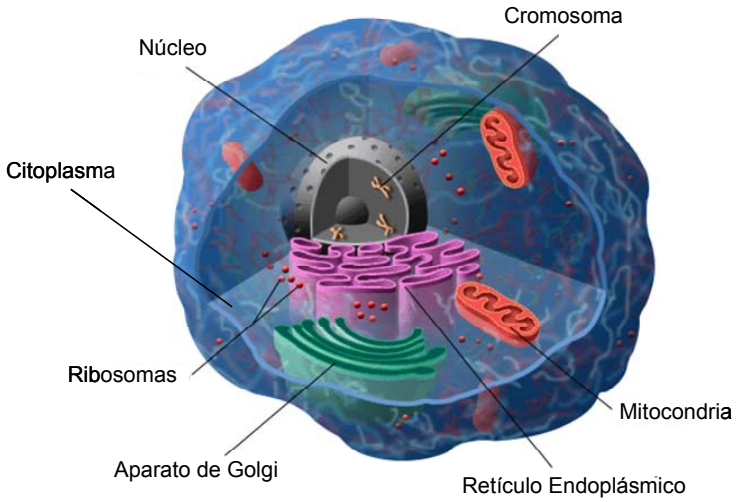


Fig. 2.1 Estructura de la Célula

Dentro de la célula se encuentran estructuras llamadas organelos como el núcleo, membrana celular, ribosomas, retículo endoplásmico, aparato de Golgi y mitocondria.

El núcleo es el centro de control de la célula y el elemento más prominente; de forma generalmente esférico, rodeado de una envoltura nuclear, en este lugar se deposita la información genética de la célula (ADN).

La membrana celular es la parte externa de la célula que envuelve el citoplasma. Permite el intercambio entre la célula y el medio que la rodea. Intercambia agua, gases y nutrientes, y elimina elementos de desecho.

El citoplasma es una suspensión coloidal, por donde se mueven o nadan los organelos dentro de la célula.

Existen diversos tamaños de células, pero en su gran mayoría son microscópicas, y sólo se pueden observar con la ayuda del microscopio, se medida es la micra, por lo cual el especialista debe contar con un

microscopio de gran precisión para poder determinar el tipo de célula, pero en muchos de los casos esto no sucede.

Para el presente trabajo es importante conocer la composición de las células para así poder determinar correctamente las características que las definen.

2.2 Composición de la Sangre

La sangre es una sustancia líquida que circula por el organismo a través del sistema circulatorio y transporta los elementos necesarios para realizar sus funciones vitales, proporciona oxígeno y nutrientes a cada célula y recoge el dióxido de carbono y las sustancias de desecho producidas por esas células [16, 48, 57].

La sangre está compuesta por un parte líquida que es el plasma y una parte sólida que son las células sanguíneas, de las cuales existen tres tipos, glóbulos rojos, glóbulos blancos y plaquetas, como se muestra en la figura 2.2 [28, 65]. El total de plasma en la sangre es del 55%, el 1% es de plaquetas, el 3% de glóbulos blancos y el 45% de glóbulos rojos.

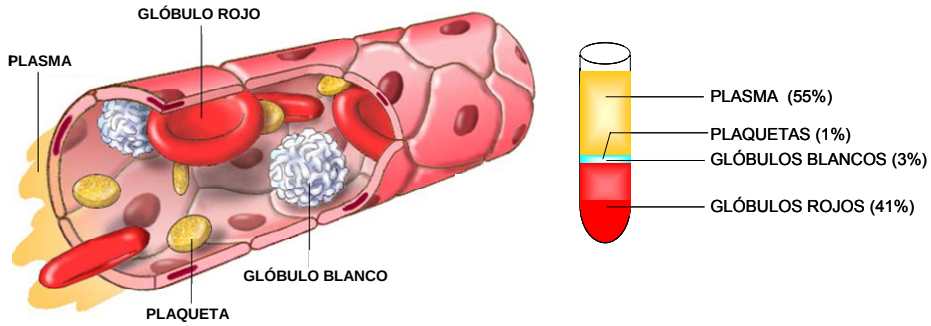


Fig. 2.2 Elementos de la Sangre

El plasma es un líquido de color amarillento, en él se encuentran las células sanguíneas. Está compuesto de agua, proteínas, sales minerales y otras sustancias necesarias para el funcionamiento normal del organismo.

Las células sanguíneas son:

- Glóbulos rojos o Hematíes, su función es transportar el oxígeno y tienen forma de discos bicóncavos aplanados. de 7 a 8 micras de diámetro
- Glóbulos blancos o Leucocitos, se encargan de proteger al organismo contra los diferentes tipos de microbios.

  De acuerdo a su apariencia en el microscopio luego de su tinción, se clasifican en granulositos que presenta gránulos en su citoplasma, con núcleo redondeado y lobulado (neutrófilos, eosinófilos y basófilos), y agranulocitos, los cuales no presenta gránulos en su núcleo (linfocitos, monocitos), como se muestra en la Figura 2.3 [10, 16].
- Plaquetas o Trombocitos, son las células sanguíneas más pequeñas e intervienen en la coagulación de la sangre.

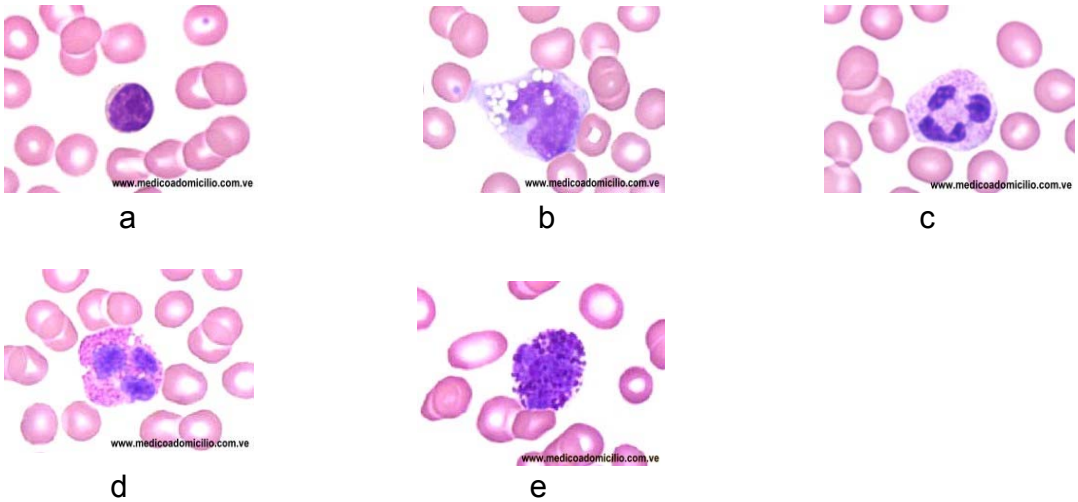


Fig. 2.3 Glóbulos Blancos o Leucocitos a) linfocito b) monocito c) neutrófilo d) eosinófilo y e) basófilo

Las células sanguíneas se producen dentro de los huesos en un espacio esponjoso llamado médula ósea, mostrado en la Figura 2.4 [29].

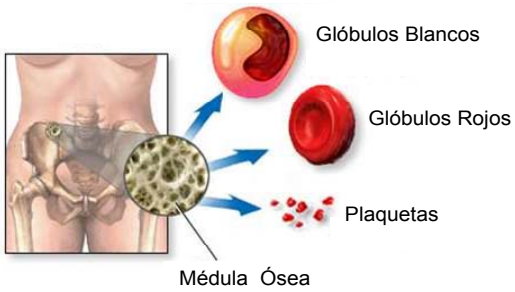


Fig. 2.4 Médula Ósea

Las células sanguíneas formadas en la médula ósea empiezan como células precursoras o células madre. La célula madre es la fase inicial de todas las células de la sangre. A medida que la célula madre madura, se desarrollan diferentes tipos de células, como los glóbulos rojos, los glóbulos blancos y las plaquetas, a este proceso se le conoce como hematopoyesis. La Figura 2.5 muestra lo anterior [46].

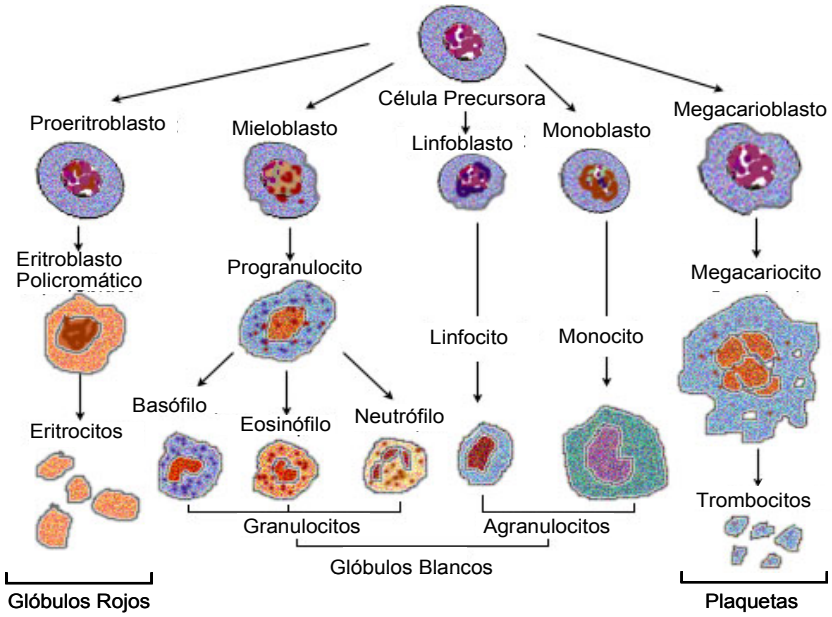


Fig. 2.5 Hematopoyesis

Las células inmaduras en la etapa temprana se llaman blastos, los cuales crecen a células maduras sanguíneas.

Los glóbulos blancos, linfocitos y monocitos son producidos en el tejido linfoide del bazo, el timo y los ganglios linfáticos, el resto de los glóbulos blancos, se producen en la médula ósea junto con los glóbulos rojos.

El tejido linfático está formado por diferentes tipos de células, la principal es llamada linfocito, su tamaño está entre 8 y 12 μm, tiene un núcleo esférico que se tiñe de violeta-azul y en su citoplasma frecuentemente se observa como un anillo periférico de color azul. Existen dos tipos de linfocitos, linfocitos B y linfocitos T.

El mieloblasto es la célula inmadura de los granulositos, pero cuando madura se generan tres tipos distintos de células: neutrófilos, eosinófilos y basófilos, éstos se diferencian por el tamaño y el color de sus gránulos [10].

Los monocitos protegen al organismo de las bacterias, tiene un tamaño de 15 a 20 μm, presenta un núcleo en forma de riñón, su citoplasma es abundante y de color gris azulado pudiendo estar acompañado de vacuolas blanquecinas.

2.3 Leucemia Aguda

Las leucemias agudas son neoplasias derivadas de las células hematopoyéticas, que proliferan inicialmente en la médula ósea antes de diseminarse a la sangre periférica, bazo, ganglios linfáticos y, finalmente a

los demás tejidos. La inmadurez de la célula que prolifera es lo que define a una leucemia como aguda.

La característica más destacada de las células neoplásicas es un defecto en la maduración más allá de la fase de mieloblasto en la LAM y de la fase de linfoblasto en la LAL. Las células leucémicas en proliferación se acumulan en la médula ósea, eliminando la hematopoyesis normal [25].

La causa precisa es desconocida, su proliferación por medio de divisiones sucesivas a partir de una célula precursora constituye el origen de las leucemias agudas, tanto linfoblásticas como mieloblásticas. Se sabe que hay diversos factores que predisponen a sufrir estas hemopatías. Entre ellos destacan los genéticos, las inmunodeficiencias, ciertos factores ambientales y los virus

Para generar la base de datos se tomaron en cuenta 1028 imágenes digitales de frotis de médula ósea de leucemias agudas, de 74 pacientes del periodo 1996 al 2006, puesto que la evaluación se realiza sobre leucemias agudas, las regiones de interés (glóbulos blancos) que se segmentarán, son las células linfoblásticas para LAL y mieloblásticas para LAM.

2.4 Clasificación de Leucemias (FAB)

La necesidad de una clasificación radica en la posibilidad de seleccionar un tratamiento más adecuado para los pacientes. La clasificación más reciente es la de la Organización Mundial de la Salud (OMS) [24] aunque la más ampliamente utilizada es la del Grupo Cooperativo Franco-Americano-Británico (FAB) [5].

La clasificación FAB [5] está basada en características morfológicas y citoquímicas y toma en cuenta el nivel de maduración en el que se encuentran los blastos y la participación de las diferentes líneas celulares. Los grupos de leucemia aguda linfoblástica y mieloblástica se subdividen en tres y seis grupos respectivamente:

La leucemia linfoblástica se divide en tres subtipos: L_1, L_2 y L_3, de acuerdo con la ocurrencia de rasgos citológicos individuales y según el grado de heterogeneidad en la población de las células leucémicas.

Tabla 2 Clasificación morfológica de Leucemia Aguda Linfoblástica según FAB.

Subtipo	Tipo(s) celular(es)	Características
LAL_1	Microlinfoblastos	Predominio de células pequeñas, con cromatina homogénea, su forma del núcleo es regular y en ocasiones hendido, los nucléolos no son visibles o son pequeños, tiene escasa cantidad de citoplasma.
LAL_2	Grandes Linfoblastos	Predominio de Células grandes y de tamaño heterogéneo, con cromatina variable y heterogénea, su forma del núcleo es irregular, los nucléolos a menudo son visibles y contiene uno o más, además tiene una cantidad moderadamente abundante de citoplasma y es variable.
LAL_3	Tipo Burkitt	Predominio de Células grandes y de tamaño homogéneo, con cromatina homogénea y un punteado fino, su forma del núcleo es regular oval o redondo, los nucléolos son muy visibles y contiene uno o más, además tiene una cantidad moderadamente abundante de citoplasma.

La descripción inicial en 1976 [5] definía seis grupos de leucemias mieloblásticas (M_1, M_2, M_3, M_4, M_5, M_6), posteriormente en 1985 [6] se le añadieron la M_7 o leucemia megacarioblástica y la M_0 o mínimamente diferenciada [7] en 1991.

Estos ocho subtipos de LAM se dividen de acuerdo a la diferencia de una o más líneas celulares y el grado de maduración.

Tabla 3 Clasificación morfológica de Leucemia Aguda Mieloblástica según FAB.

Subtipo	Tipo(s) celular(es)	Características
LAM_0	Mieloblastos	Los blastos son de citoplasma abundante sin granulación y núcleo de cromatina laxa con 1-2 nucleolos, y es mínimamente diferenciada.
LAM_1	Mieloblastos	Mieloblastos no granulares que suelen contener uno o más nucleolos nítidos. Son de núcleo redondo, el citoplasma sin granulación o con fina granulación incipiente apenas perceptible en microscopía óptica.
LAM_2	Mieloblastos, promielocitos, mielocitos	Maduración más allá de la etapa del promielocito. Son comunes a las células que contienen bastones de Auer.
LAM_3	Promielocitos hipergranulares	La mayoría de las células son promielocitos con un patrón característico de intensa granulación.
LAM_4	Promielocitos, mielocitos, promonocitos, monocitos	Existe diferenciación granulocítica y monocítica en proporciones variables.
LAM_5 (a)	Monoblastos	Monoblásticas poco diferenciadas. Los blastos son grandes, de núcleo redondo con tendencia a situarse centralmente en el citoplasma amplio.
LAM_5 (b)	Monoblastos, promonocitos, monocitos	Monoblastos, promonocitos y monocitos diferenciados. Morfológicamente los blastos presentan núcleos arriñonados con frecuentes plegamientos y el citoplasma es azul basófilo con discreta granulación.
LAM_6	Eritroblastos	Eritropoyesis grotesca. Eritroblastos con lobulación múltiple del núcleo, núcleo múltiples, fragmentos nucleares, formas gigantes y rasgos megaloblásticos.
LAM_7	Megacariocitos	Leucemia megacarioblástica aguda en la cual se destacan megacariocitos inmaduros con plaquetas anormales. Morfológicamente son de aspecto heterogéneo, pueden presentar mamelones citoplásmicos y vacuolización o aspecto seudolinfoide o seudomonocitoide.

La base de datos esta compuesta de imágenes de frotis de médula ósea, solo con los subtipos L_1, L_2, M_2, M_3 y M_5, que fueron las únicas muestras que nos proporcionó el IMSS, San José.

Utilizando las diferencias morfológicas según la clasificación FAB, nos permitieron obtener de los subtipos características de tipo geométrico, de textura y estadísticas para su posterior clasificación.

2.5 Diagnostico de Leucemias

La leucemia es un cáncer de las células sanguíneas. Existen diferentes subtipos de leucemias, de acuerdo con las células sanguíneas afectadas. Cada leucemia presenta características distintas de la enfermedad y por lo tanto, diferentes opciones de tratamiento. Se están utilizando diversas pruebas de diagnóstico clínico con el fin de determinar el subtipo y tipo de leucemia.

Actualmente no hay pruebas especiales para la detección de la leucemia. La mejor estrategia para un diagnóstico temprano es tomar en cuenta los signos y síntomas que presenta el paciente, pero es necesario realizar una serie de análisis clínicos, que detecten la presencia de las células anormales.

El examen físico inicia con un estudio del cuerpo para comprobar los signos generales de salud, inclusive el control de síntomas de enfermedad, como masas o cualquier otra cosa que parezca anormal. Se toman también en cuenta los antecedentes médicos de las enfermedades y los tratamientos previos del paciente.

Los análisis de sangre ayudan al médico a diagnosticar y controlar una enfermedad. Aparte de examinar las células sanguíneas, existen diversas sustancias químicas en la sangre que dan información importante acerca del funcionamiento de los sistemas del organismo.

Un recuento sanguíneo completo es una prueba para contabilizar el número de glóbulos rojos, glóbulos blancos y plaquetas en una muestra de sangre (periférica o de médula ósea), donde se mide su tamaño, número y madurez en los diferentes tipos de células sanguíneas [11, 31] y se emplea para determinar muchos tipos de anomalías. Cualquier variación en la cantidad, tamaño o madurez de los parámetros normales de las células sanguíneas, son utilizados para indicar una infección o enfermedad. También con una parte de la muestra de sangre se hace un frotis, extendiendo una pequeña gota de sangre sobre una lámina de vidrio para formar una película delgada, se deja secar y se le aplican ciertos tintes. Los tintes colorean los diferentes tipos de células sanguíneas de manera que se puedan distinguir unas de las otras. Se examina la lámina con un microscopio, se cuentan los distintos tipos de glóbulos blancos y se observan las células para determinar si son normales o anormales. El proceso anterior se muestra en la figura 2.6

Fig. 2.6 Análisis de Sangre Periférica

Algunas formas de cáncer pueden afectar la producción de células sanguíneas en la médula ósea. Un aumento en la cantidad de glóbulos

blancos inmaduros en un recuento sanguíneo completo puede estar asociado con la leucemia.

El recuento de estas células se usa durante el diagnóstico, tratamiento y seguimiento para determinar la salud del paciente. El recuento de células sanguíneas por sí solo no determina, si el paciente tiene un cáncer relacionado con la sangre, pero puede advertir al médico de que se requieren más pruebas, por ejemplo un estudio de médula ósea.

La aspiración de médula ósea es otro análisis donde se extrae una pequeña porción de este tejido para su análisis [4, 26]. El médico toma la muestra de médula ósea del hueso ilíaco o del esternón, aunque ocasionalmente se puede seleccionar otro hueso. Una parte de la muestra queda en el tubo para su contabilización y otra parte de la muestra se coloca en una lámina de vidrio para ser examinado bajo el microscopio. Este proceso se muestra en la figura 2.7

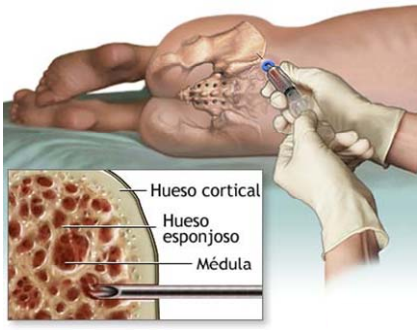


Fig. 2.7 Aspiración de Médula Ósea

En la médula ósea también existe una parte sólida, una biopsia de médula ósea es tomar una pequeña porción de la parte sólida. Para realizar una biopsia de hueso se hace una pequeña incisión en la piel, mostrada en la figura 2.8.

Fig. 2.8 Biopsia de Médula Ósea

La citometría (Cito=célula, metría=medición) es el análisis de las características particulares de las células, mediante el microscopio o de manera automatizada.

Otro tipo de análisis es la citometría de flujo (CMF) es una técnica que puede medir diversos parámetros en decenas de millares de células individuales en pocos segundos. Sus aplicaciones son numerosas y se emplean tanto en la investigación biológica como médica, las más importantes se relacionan con la hematología e inmunología clínicas, en la que se miden parámetros de número y clasificación de células sanguíneas.

Esta técnica también se utiliza para la clasificación de leucemias agudas, entre otros padecimientos [3]. Su principal característica es que es multiparamétrica, por lo que proporciona información simultánea de varios parámetros de cada una de las células analizadas y la relación entre los parámetros de las células examinadas. Se basa en hacer pasar las células alineadas de una en una por un haz de luz láser que produce señales que corresponden a diferentes parámetros de la célula y que son recogidos por distintos detectores [14]. Como se muestra en al figura 2.9 [3].

Los parámetros son:

- Los relacionados con características específicas de la célula, como tamaño y complejidad de su núcleo y citoplasma.
- Los relacionados con características antigénicas de cada célula (inmunofenotipo).

Por esto la CMF puede identificar una célula por medio de sus características antigénicas y/o por sus características morfológicas de tamaño y complejidad y se aplica a muestras de sangre que pueden proceder de Médula ósea o periférica [45].

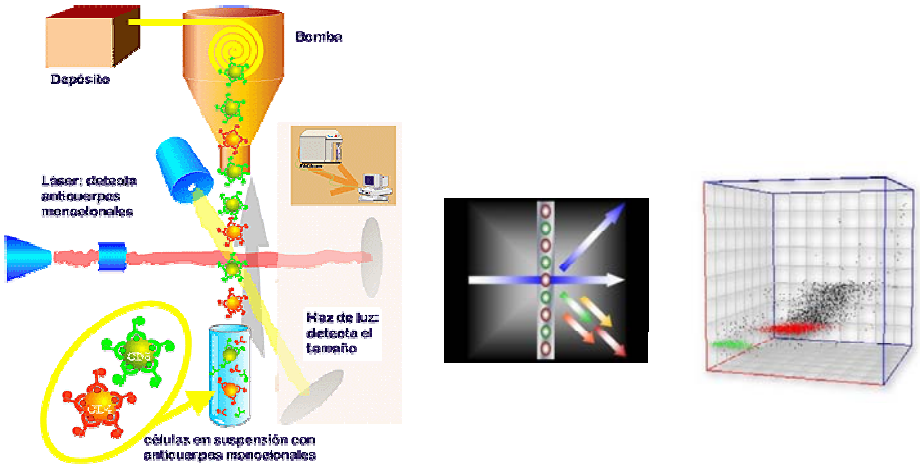


Fig. 2.9 Citometría de Flujo

Utilizando algunos de los análisis anteriores los especialistas del IMSS San José detectaron que las muestras de 74 pacientes tenían leucemia.

Como primer análisis les realizó una aspiración de médula ósea para hacer un conteo de células y un estudio morfológico del frotis de sangre y se detectó que las muestras tenían leucemia. Para comparar los resultados anteriores se realizo otro análisis, una citometría de flujo con inmunofenotipo para poder determinar el tipo y subtipo de leucemia que tenía cada paciente, por lo que el estudio morfológico se complemento con la citometría de flujo.

Los estudios morfológicos hasta el momento no dan información tan precisa (por la forma manual de realizarlos) para el diagnóstico de leucemia como la citometría de flujo, pero se utilizan como estudio inicial para determinar si tiene o no leucemia un paciente.

Por lo que el presente trabajo se enfoca en el estudio de la morfología de los glóbulos blancos.

Capítulo 3

Visión por Computadora para la obtención de Regiones de Interés

En este capítulo se describen algunos conceptos básicos de visión por computadora, específicamente los conceptos que se utilizan en el método propuesto, como es la segmentación de regiones de interés y la extracción de características de éstas. La sección 3.1 detalla en que consiste el procesamiento de imágenes. La sección 3.2 presenta la segmentación de regiones de interés por color y por detección de bordes. La sección 3.3 describe la obtención de características. La sección 3.4 muestra la clasificación de características. La sección 3.5 explica la computación en imágenes médicas. Por último la sección 3.6 muestra el estado de arte.

3.1 Procesamiento de Imágenes

La visión es la ventana al mundo de muchos organismos. Su función principal es reconocer y localizar objetos en el ambiente mediante el procesamiento de las imágenes [67].

El objetivo de la visión por computadora (VC) es tomar decisiones útiles acerca de los objetos físicos reales del mundo (de escenas) con base a imágenes adquiridas digitalmente.

Por lo tanto, la tarea de la VC es la construcción de descriptores de la escena con base a características relevantes contenidas en una imagen [54].

Una imagen monocromática puede ser definida como una función bidimensional *f(x,y)*, donde *x* y *y* son coordenadas espaciales, y el valor de *f* en cualquier par de coordenadas se denomina intensidad o nivel de gris de la imagen en el punto (mostrado en la figura 3.1).

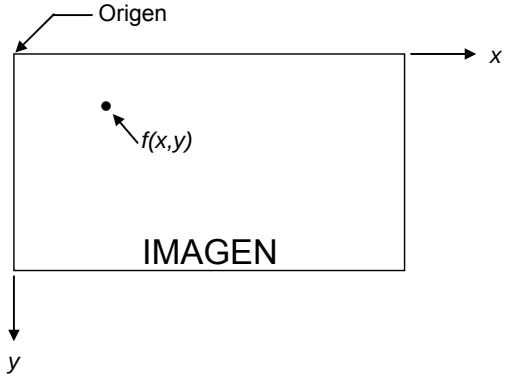


Fig. 3.1 Representación de una Imagen

Una imagen digital puede considerarse como una matriz cuyos índices del renglón y columna identifican un punto en la imagen y el correspondiente valor del elemento de la matriz es el que identifica el nivel de intensidad de luz en ese punto [22]. Los elementos del arreglo digital se llaman elementos de imagen o píxeles (abreviatura del inglés picture elements) [33]. Un Píxel es el término más comúnmente usado para denotar los elementos de una imagen digital.

Para trabajar con los valores de cada píxel en la computadora (nivel de brillo) se utilizan números enteros positivos binarios. El número de niveles está determinado por la siguiente relación (1).

$$L = 2^{B} \tag{1}$$

Donde B es el número de niveles de brillo que tiene la imagen. Para una apreciación fina se utilizan 8 bits (256 niveles de gris) [35]. Una imagen binaria es una imagen monocromática que tiene 1 BIT por píxel, esto es, dos niveles, blanco ó negro.

En procesamiento digital de imágenes se utilizan un conjunto de técnicas que trabajan sobre la representación digital de una imagen, cuyo objetivo es obtener algunos de los elementos que conforman la imagen, de modo que sea más fácil su análisis posterior, por parte de un sistema de visión artificial o un usuario.

Los principales pasos en el procesamiento de imágenes se muestran en la siguiente figura 3.2 [29]

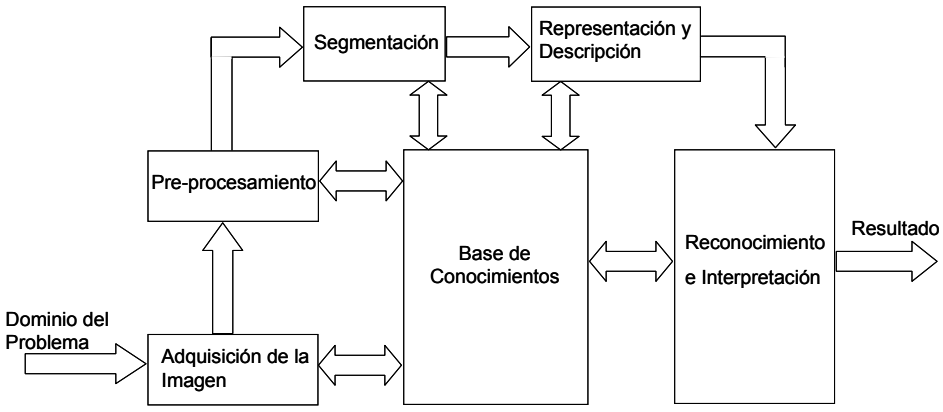


Fig. 3.2 Modelo General para el Procesamiento de Imágenes

De acuerdo al Modelo General para el Procesamiento de Imágenes, este trabajo se enfoca a solo tres procesos, la segmentación, la descripción y el reconocimiento.

3.2 Proceso de Segmentación de las Regiones de Interés

El primer paso que se utilizó del modelo general (figura 3.2) para procesamiento de imágenes fue la segmentación. La segmentación de una imagen consiste en la división o separación de la misma en regiones con atributos similares [56].

El objetivo de la segmentación es dividir una imagen en sus regiones constituyentes u objetos. Los niveles de división de la figura, dependen del problema que va a resolver. Generalmente, la segmentación automática es una de las tareas más difíciles en el procesamiento de imágenes. La precisión de la segmentación determina el éxito eventual o falla del procedimiento de análisis automatizado.

Los algoritmos de segmentación de imágenes monocromáticas se basan generalmente en tres propiedades básicas de los valores de nivel de gris:

- Discontinuidad: este enfoque particiona la imagen con base a cambios bruscos de intensidad en los niveles de gris para poder detectar puntos, líneas y bordes. Para su detección tenemos los operadores de la primera derivada, la segunda derivada y los morfológicos.
- Similitud: la similitud en los tonos de gris de los píxeles de un entorno, de acuerdo a un conjunto de criterios predefinidos, permiten construir regiones por división y fusión, por crecimiento o por umbralización, similitud de textura, color o nivel de gris.
- Conectividad de los píxeles. Una región se dice conexa o conectada si para cada par de píxeles de la región, existe un camino de uno al

otro formado por píxeles de la misma. Un camino de píxeles es una secuencia de píxeles adyacentes. La conectividad de los píxeles desempeña un papel importante en la segmentación de imágenes. Para su detección tenemos los algoritmos de K- medias, ISODATA y máximos y mínimos.

3.2.1 Detección de Discontinuidades

Los tres tipos básicos de discontinuidad de una imagen digital son: puntos, líneas y bordes. Para el presente trabajo solo utilizaremos la detección de bordes.

El borde es la frontera entre dos regiones con propiedades de nivel de gris relativamente distintas [22, 23, 50].

El borde de una región es el conjunto de píxeles en la región que tienen uno o más vecinos que no están en la región. Píxeles que no están en la región o en el borde son llamados píxeles de fondo [22, 23, 50].

Aunque existen distintas técnicas de detección de bordes, en este trabajo se utilizó el algoritmo de trazado sencillo de borde interno [17] que a continuación se explica (ver figura 3.3).

Trazado sencillo de borde interno

- Si un borde de región no es conocido pero se han definido las regiones en la imagen, pueden detectarse únicamente los bordes.
- Asumimos que la imagen con regiones está en formato binario o que las regiones han sido etiquetadas
- Un borde interior de la región es un subconjunto de la región mientras que un borde exterior no es un subconjunto de la región.
- El objetivo es determinar los bordes interiores de la región.
- El algoritmo trazado simple de borde interior abarca la localización de los bordes con vecindad 4 y vecindad 8.

- El algoritmo trabaja para todas las regiones más grandes que un píxel.
- Este algoritmo es capaz de encontrar el borde de una región, pero no encuentra bordes de una región con agujeros.

Algoritmo: Trazado de borde interno [17]

1. Buscar en la imagen a partir de la esquina superior izquierda hasta que un píxel de una nueva región es encontrado. Este píxel P_0 entonces tiene el valor mínimo de la columna y de la fila de todos los píxeles de esa región. El píxel P_0 es un píxel de inicio del borde de la región.

   Define una variable *dir* que almacena la dirección del movimiento previo a lo largo del borde desde el elemento previo del borde hasta el elemento actual del borde. Asigna

   a) $dir = 3$ Si el borde es detectado con vecindad 4;

   b) $dir = 7$ Si el borde es detectado con vecindad 8;

2. Buscar la vecindad de 3×3 del píxel actual en la dirección de sentido contrario a la manecillas del reloj, iniciando la búsqueda del vecindario en el píxel posicionado en la dirección

   a) $(dir + 3) \mod 4$ Si el borde es detectado con vecindad 4;

   b) $(dir + 7) \mod 8$ Si dir es non;

   $(dir + 6) \mod 8$ Si dir es par.

   El primer píxel encontrado con el mismo valor que el píxel actual es un nuevo elemento borde P_n.

   Actualiza el valor dir.

3. Si el elemento actual borde P_n es igual al segundo elemento borde P_1, y si el elemento borde previo P_{n-1} es igual a P_0, parar. De lo contrario repita el paso 2.
4. El borde interior detectado está representada por píxeles $P_0 \ldots P_{n-2}$.

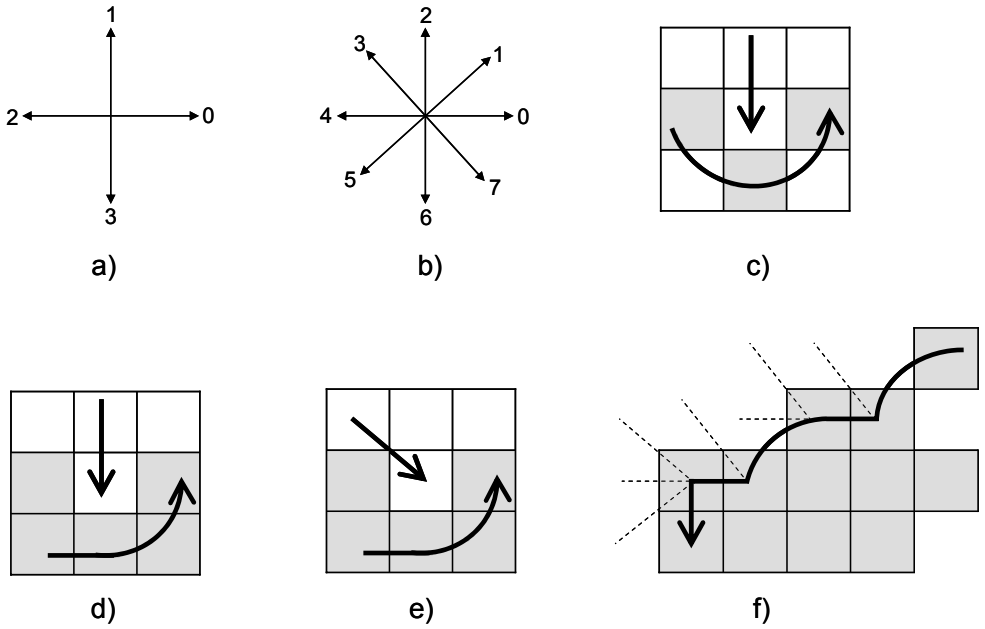


Fig. 3.3 Trazado de borde: a) Notación de dirección, vecindad 4, b) Vecindad 8, c) Búsqueda de secuencia del píxel en el vecindario con conectividad 4, d) y e) Búsqueda de secuencia de vecindad 8, f) Trazado de borde de vecindad 8.

3.2.2 Segmentación por Regiones

El objetivo de la segmentación por regiones [62] es dividir una imagen en diferentes regiones (píxeles conectados entre si), cada una de los cuales presentan propiedades o características comunes dentro de la misma imagen. Algunas de sus propiedades son nivel de gris, valor(es) de color, gradiente local, movimiento, etc.

Sea R la representación de la región completa de una imagen. Se puede contemplar la segmentación como un proceso que divide a R en n subregiones, $R_1, R_2, ..., R_n$, de tal forma que [23]:

a) $\bigcup_{i=1}^{n} R_i = R,$

b) R_i es una región conexa, $i = 1,2,...,n,$

c) $R_i \cap R_j = \phi$ para todo i y $j, i \neq j,$

d) $P(R_i) =$ VERDADERO para $i = 1,2,...,n,$ y

e) $P(R_i \cup R_j) =$ FALSO para $i \neq j,$

donde $P(R_i)$ es un predicado lógico sobre los puntos del conjunto R_i y ϕ es el conjunto vacío.

Aunque existen diversas técnicas de segmentación como umbral, regiones y agrupación, este trabajo utiliza la segmentación por agrupación basada en el color, para poder eliminar el fondo de la imagen digital y solo obtener las células de los glóbulos blancos para su posterior análisis.

3.2.3 Segmentación por Agrupación

Las técnicas de agrupamiento clasifican los píxeles estadísticamente, sin tener tomar en cuenta su información de regiones o de bordes, sólo la información de intensidad de cada punto.

El proceso de agrupación (Clustering) consiste en la división de los datos en grupos de objetos similares. Para medir la similaridad entre objetos se suelen utilizar diferentes formas de distancia: Euclidiana, de Manhattan, de Minkowski, etc [68].

Considerando entonces n píxeles y p variables, la matriz de datos es la siguiente:

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

Donde $n = 600$ y $p = 800$ que es el tamaño de imágenes digitales que contiene nuestra base de datos.

La matriz de disimilaridad sería:

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix}$$

Donde $d(j,i)$ representa la diferencia entre el píxel i y el píxel j (cuanto más pequeño sea el valor, más similares serán), y puede calcularse con la siguiente fórmula de la distancia Euclidiana, que es la se utiliza en el método propuesto:

$$d(i,j) = \sqrt{\left(|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \ldots + |x_{ip} - x_{jp}|^2\right)} \qquad (2)$$

La distancia Euclidiana es la distancia de una línea recta entre dos píxeles, como se muestra en la figura 3.4 y la ecuación 2.

0	0	0
0	1	0
0	0	0

a)

1.41	1.0	1.41
1.0	0.0	1.0
1.41	1.0	1.41

b)

Fig. 3.4 Distancia Euclidiana a) Imagen, b) Transformada de la distancia

Para el problema de segmentación también podemos utilizar técnicas de agrupamiento (Clustering). La idea básica consiste en suponer que los píxeles de la imagen constituyen puntos de un espacio de 3 dimensiones (RGB), de tal forma que los puntos de una región de color similar aparecen en este espacio como racimos agrupados en una zona muy densa. Todos los

colores característicos de la imagen tendrán pues grupos densos de puntos. Las técnicas de agrupamiento permiten obtener un punto central o representativo de cada agrupación densa de puntos, basándose en medidas de similitud o distancia entre dichos puntos. Aunque existen muchas de las técnicas clásicas de agrupamiento la técnica de K-medias [50, 68] la que se ha utilizado en este trabajo.

A continuación se describe el Algoritmo de K-medias [2, 50, 62].

Algoritmo K-medias [17, 39]

Este algoritmo parte del conocimiento a priori del número de clases pero no de sus características.

1. De entre la serie a clasificar se escogen de forma arbitraria, tantos elementos como número de clases y se considera que constituyen los centroides de cada clase.

2. El resto de los elementos se asignan a cada clase siguiendo el criterio de mínima distancia a los centroides antes elegidos.

3. Se recalculan los centroides de cada clase. Para ello se toma la media de todos los valores dentro de cada clase.

4. Se vuelven a asignar, ahora todos los elementos, a cada clase con el criterio de mínima distancia.

5. Se vuelven a calcular los centroides. Si no varían se considera que el algoritmo ha terminado, si no, se vuelve a repetir el paso anterior.

3.2.4 Modelos de Color

El uso del color en el procesamiento de imágenes está motivado por porque el análisis de imágenes el color es un potente descriptor que a menudo simplifica la identificación y extracción de objetos en una imagen.

Modelo RGB

Las imágenes de nuestra base de datos están en formato RGB. La descripción RGB (del inglés Red, Green, Blue) de un color hace referencia a la composición del color en términos de la intensidad de los colores primarios con que se forma: el rojo, el verde y el azul. Es un modelo de color basado en la síntesis aditiva, con el que es posible representar un color mediante la mezcla por adición de los tres colores luz primarios [66]. El modelo de color RBG se muestra en la figura 3.5.

Fig. 3.5 Modelo de Color RGB

Modelo CIELAB

La CIE (Comisión Internacional de l'Eclairage) propuso el modelo CIE Lab, el cual dimensiona la totalidad del espectro visible.

Los tres colores de luz percibidos RGB se miden en el contexto de una iluminación específica y todos los demás son considerados como una combinación de color, iluminación y superficie reflectante. En cambio Lab considera el espacio en forma uniforme y despliega tres ejes espaciales: *L* (luz, blanco-negro), *a* (rojo-verde), *b* (amarillo-azul).

El componente de luminosidad (*L*) oscila sus valores entre un rango de 0 y 100. El componente *a* (eje verde-rojo) y el componente *b* (eje azul-amarillo) pueden estar comprendidos entre + 120 y - 120. El modo *Lab* se

usa sobre todo cuando se desea modificar los valores de luminosidad y color de una imagen por separado.

En este trabajo además de utilizar el modelo de color RGB, también utilizamos el modelo de color CIELab [44] (ver figura 3.6), porque para hacer el proceso de segmentación por agrupación solo se trabaja con una matriz que tiene los componentes *a* y *b* (de CEILab) y no con tres del formato RGB, además como utiliza todo el espectro visible se puede diferenciar mejor los glóbulos blancos del resto de los componentes de la imagen digital.

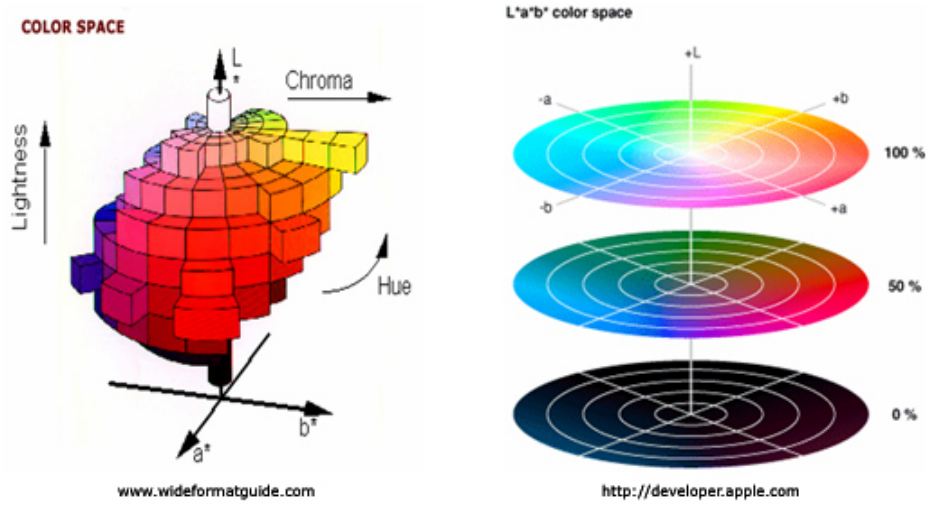


Fig. 3.6 Modelo de Color CIEL*a*b*

3.3 Proceso de Descripción de las Regiones de Interés

El segundo paso que se utilizó del modelo general para procesamiento de imágenes (figura 3.2) fue el proceso de descripción. Y es también conocida como selección de características [23], y trata de extraer los rasgos o características que son básicas para diferenciar una región de otra.

Una región puede describirse por la forma de su frontera o por sus características internas. Para clasificar las diferentes regiones, extraemos de ellas una serie de características internas. Podemos obtener características de propiedades topológicas, geométricas, estadísticas y de textura.

Una región de interés (ROI) puede ser vista como un conjunto de puntos conectados entre si.

Una propiedad topológica se caracteriza por ser invariante a escala, rotación y traslación. Algunos de los descriptores topológicos son número de huecos, número de componentes conectados y el número de Euler (Ver fórmula 3).

$$E = C - H \tag{3}$$

Donde E es el número de Euler, C es el número de componentes conectadas y H el número de huecos.

Otras propiedades son las geométricas entre las que destacan el área, perímetro y el centro de gravedad. El perímetro es la longitud de su frontera. El área es el número de píxeles contenidos dentro de su frontera, como se muestra en la siguiente fórmula.

$$A = \sum_{i=1}^{n} \sum_{j=1}^{m} x_{ij} \tag{4}$$

Existen otro tipo de descripciones como las de textura cuya característica principal es la repetición de patrón básico. También están las descripciones estadísticas, entre éstas tenemos: la media (fórmula 5), la varianza (fórmula 6), la desviación estándar (fórmula 7) y el análisis de componentes principales (ACP) [34] entre otros.

$$\overline{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i \tag{5}$$

$$s^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2 \tag{6}$$

$$s = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2} \tag{7}$$

El Análisis de Componentes Principales (ACP) es una técnica estadística de síntesis de información, o reducción de la dimensión del número de variables. El objetivo es reducir el número de variables perdiendo la menor cantidad posible de información. Los nuevos componentes principales o factores serán una combinación lineal de las variables originales, y además serán independientes entre sí.

Un análisis de componentes principales tiene sentido si existen altas correlaciones entre las variables, ya que esto es indicativo de que existe información redundante y, por tanto, pocos factores explicarán gran parte de la variabilidad total.

La elección de los factores se realiza de forma tal que el primero factor recoja la mayor proporción posible de la variabilidad original; el segundo factor debe recoger la máxima variabilidad posible no recogida por el primero, y así sucesivamente. Del total de factores se elegirán aquéllos que recojan el porcentaje de variabilidad que se considere suficiente. A éstos se les denominará componentes principales.

Para realizar el ACP partimos de una matriz X de dimensiones $n \times p$, donde p corresponde al número de variables. Después la matriz X, debe ser estandarizada mediante la siguiente fórmula

$$y_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j} \tag{8}$$

donde $\bar{x}_j$ y s_j son el promedio y la desviación estándar para cada una de las variables.

El siguiente paso es obtener la matriz de correlación de la matriz tipificada utilizando la siguiente fórmula

$$S = \frac{COV(x)_{i,j}}{\sqrt{VAR(x)_i} * \sqrt{VAR(x)_j}} \tag{9}$$

Y por último para obtener los eigenvalores resolvemos la ecuación característica, donde I es la matriz identidad, S es la matriz de correlaciones.

$$|S - lI| = 0 \tag{10}$$

La ecuación matricial del ACP es:

$$L = U'SU \tag{11}$$

Donde L es la matriz de componentes principales, S es la matriz de correlación de la matriz tipificada y U la matriz de vectores propios de la matriz S.

Los elementos de la diagonal de L, $l_1, l_2, \ldots, l_p$ son llamados valores propios (eigenvalores) de S. Las columnas de U, $u_1, u_2, \ldots, u_p$ son llamados vectores propios (eigenvectores) de S.

En nuestro proceso obtuvimos diferentes características geométricas, estadísticas, de textura y topológicas. Una vez obtenidas las características

descriptivas de las regiones de interés las utilizamos para el proceso de clasificación, como se describe en la siguiente sección.

3.4 Proceso de Clasificación de las Regiones de Interés

El último paso que se utilizó del modelo general para procesamiento de imágenes (figura 3.2) fue la clasificación y en este proceso es donde se realiza el reconocimiento e interpretación de las ROI's.

El reconocimiento es el proceso que etiqueta, o asigna un nombre, a un objeto basándose en la información que proveen sus descriptores. La interpretación involucra la asignación de significado a un conjunto de objeto reconocido.

El objetivo principal de las técnicas de reconocimiento de ROI's, aplicadas a un problema general de clasificación consiste en asignar a una ROI una de las diversas clases previamente especificadas.

En el reconocimiento de patrones existen dos tipos de clasificación la supervisada y la no supervisada [17]. La diferencia fundamental entre ambos métodos estriba en si se conoce o no la clase a la cual pertenece cada patrón.

La clasificación no supervisada enfoca la clasificación como el descubrimiento de las clases del problema. Los objetos únicamente vienen descritos por un vector de características, sin que sepamos la clase a la que pertenece cada uno de ellos.

La clasificación supervisada parte de un conjunto de objetos descritos por un vector de características y la clase a la que pertenece cada uno de ellos. A este conjunto de objetos se le denomina *conjunto de entrenamiento.*

Así que tomando como base el *conjunto de entrenamiento* la clasificación supervisada construye un *modelo* que se utilizará para clasificar nuevos objetos de los cuáles no conozcamos su clase.

La clasificación supervisada ha sido utilizada en numerosos problemas de distinta índole como el diagnostico de enfermedades, la concesión o rechazo de créditos bancarios, predicción de bancarrota en empresas, reconocimiento de caracteres escritos a mano, detección de anormalidades en cromosomas, etc., y es el tipo de clasificación que utiliza nuestro modelo propuesto.

3.4.1 Métodos de Clasificación Supervisada

- **Vecino más próximo (1-NN):** el clasificador asociará al caso x la clase verdadera del objeto que se encuentra más próximo a x dentro del espacio de representación. Utiliza la distancia Euclidiana para encontrar la instancia de entrenamiento más cercana a la instancia de prueba dada, y predecir la misma clase que esta instancia de entrenamiento.

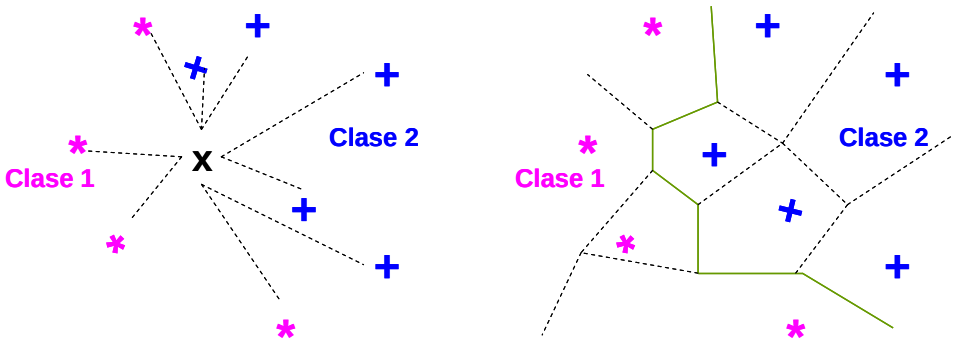


Fig. 3.7 a) Clasificación 1-NN. b) En trazo continuo, la frontera de decisión; en trazo discontinuo, los bordes de la partición de Voronoi asociada.

- **Vecinos más próximos (K-NN):** la clase asignada a un nuevo caso x será la clase más votada entre los K vecinos más próximos del conjunto de entrenamiento.

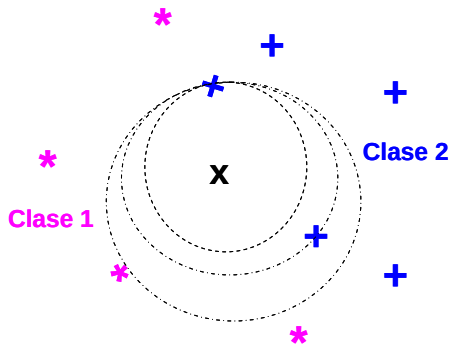


Fig. 3.8 Clasificación 3-NN.

- **Kstar (K*):** es un clasificador basado en instancias, esto significa que la clasificación de una instancia está basada en la clasificación de instancias de entrenamiento similares, determinadas por alguna función de similitud. Usa una función de distancia basada en entropía.
- **Árboles de clasificación:** Todo árbol de clasificación comienza con un nodo al que pertenecen todos los casos de la muestra que se quiere clasificar. Se le denomina nodo *raíz*, el resto de los nodos se dividen en nodos *intermedios* o no terminales, y nodos *hoja* o terminales es decir nodos que no se van a dividir más. En la fase de construcción del árbol cada nodo *hoja* se hace corresponder con una categoría completa de la variable clase. De esta manera los nodos hojas representan las diferentes particiones en las que se ha dividido el espacio de clasificación.

  A la hora de clasificar cada patrón, el punto de partida es el nodo *raíz*, dependiendo de los valores de la variable predictora por la que se pregunta, los casos se van distribuyendo por los nodos

hijos. El proceso se repite en cada nodo hasta llegar a los nodos *hoja*.

El algoritmo **C4.5** se basa en la utilización del criterio de ganancia de información (gain ratio), De esta manera se consigue evitar que las variables con mayor numero de posibles valores salgan beneficiadas en la selección. También incorpora una poda del árbol de clasificación una vez que este ha sido inducido y esta basada en la aplicación de una prueba de hipótesis que trata de responder a la pregunta de si merece la pena expandir o no una determinada rama

Para el algoritmo **RandomForest** en cada nodo se seleccionan de forma aleatoria algunas de las variables de entre todas las disponibles. La decisión se realizará en función de las variables seleccionadas. Se basan en el desarrollo de muchos árboles de clasificación. Para clasificar un nuevo objeto desde un vector de entrada, ponemos dicho vector bajo cada uno de los árboles del bosque. Cada árbol genera una clasificación, es decir cada árbol vota por una clase. El bosque escoge la clasificación teniendo en cuenta el árbol más votado sobre todos los del bosque. Cada árbol crece de la forma más extensa posible, sin ningún tipo de poda.

Un modelo de árbol logístico (**LMT**) básicamente consiste de una estructura de árbol de decisión estándar con funciones de regresión logísticas en las hojas, más que un modelo de árbol es un árbol de regresión con funciones de regresión en las hojas. Como en árboles de decisión ordinarios, una prueba sobre uno de los atributos esta asociada con cada nodo interior. Para atributos numéricos, el nodo tiene dos nodos hijos y la prueba consiste en comparar el valor de atributo a un umbral: una instancia es colocada debajo de la rama izquierda si su valor para aquel atributo

es más pequeño que el umbral y colocada debajo de la rama derecha en caso contrario (ramificaciones binarias).

ADTree: Es una aplicación del método de amplificación (*boosting*) a los árboles de decisión. Representa los árboles en una estructura alterna que permite ver cada nodo como un tronco. Un tronco (*stump*) es un árbol truncado que sólo contiene una bifurcación. Las dos ramas de cada tronco son los nodos de predicción, que contienen un valor numérico. Estos valores son utilizados luego en una votación en el momento de clasificar una nueva instancia.

- **NaiveBayes:** Este algoritmo se basa en la hipótesis de que las variables que describen a las instancias son estadísticamente independientes. A partir del conjunto de entrenamiento se calcula la probabilidad a priori de que una instancia cualquiera pertenezca a una clase, también se calcula la probabilidad condicional de que un atributo tome un valor si la instancia pertenece a una determinada clase, luego con estos datos se puede calcular utilizando la fórmula de Bayes y asumiendo independencia entre las variables, la probabilidad de que una instancia pertenezca a una clase si sus atributos toman determinados valores. La clasificación de la instancia dada será la que haga máxima esta probabilidad.

- **SMO:** El algoritmo SMO es un método rápido para entrenar máquinas SVMs. El entrenamiento de un SVM requiere la solución a un gran problema de optimización de programación cuadrática. SMO divide este problema en una serie de problemas más pequeños que se resuelven de forma analítica.

  Se trata de un tipo de red caracterizado por tener un aprendizaje no supervisado competitivo y una salida bidimensional. A partir de

un patrón de entrada, cada neurona de la capa de salida calcula la similitud entre su propio vector de pesos y el vector de entrada según una medida de distancia establecida. Se considera vencedora a la neurona cuya similitud sea mayor**.**

AdaBoost (Adapting Boosting) generan igual un conjunto de clasificadores secuencialmente (Bagging los puede generar en paralelo). Combina la decisión de los clasificadores por votos ponderados, es decir a todos los ejemplos, les asigna inicialmente un peso igual ($1/m$). El nuevo clasificador que se cree debe prestar más atención a aquellos ejemplos en los que los anteriores han producido errores

Cada vez que se genera un clasificador, se cambian los pesos de los nuevos ejemplos usados para el siguiente clasificador. La idea es forzar al nuevo clasificador a minimizar el error esperado. Para esto se les asigna más peso a los ejemplos mal clasificados y menos a los bien clasificados [69, 70].

Los procedimientos de clasificación supervisada explicados anteriormente fueron utilizados en nuestro método propuesto.

3.5 Computación en Imágenes Médicas

En tiempos recientes el desarrollo de las computadoras ha permitido un aumento en el uso de las imágenes digitales en el campo informático, dada la diversidad y cantidad de información que poseen.

El procesamiento de imágenes médicas se ha convertido en un campo importante en la Visión Artificial. El objetivo principal es la obtención de información para ayudar al médico a establecer un diagnóstico, elegir o controlar un tratamiento.

A pesar de que estas imágenes pueden ser estudiadas en el microscopio por el especialista y proveer información respecto ala cantidad y morfología de las células, la interpretación puede ser un poco subjetiva si no se cuenta con el conocimiento, experiencia e instrumentos necesarios. Debido a esto, el uso de la computadora puede ayudar a extraer información precisa, repetible y a encontrar patrones que nos ayuden a determinar el tipo de región de interés de que se trata.

Este aspecto constituye nuestra área de investigación. Se trata de construir una nueva herramienta de apoyo al diagnóstico médico, basada en un análisis semi-automático de imágenes médicas.

El tratamiento digital de imágenes contempla el procesamiento y el análisis de las imágenes.

El objetivo del procesamiento de imágenes es mejorar su calidad mediante transformaciones, restauraciones y mejoramiento para su posterior utilización o interpretación. El análisis de imágenes consiste en la extracción de propiedades y características de las imágenes, así como la clasificación e identificación y el reconocimiento de patrones.

3.5 Estado del Arte

Existen varios grupos de investigación que han creando diversos métodos para la segmentación y clasificación de células sanguíneas, ya sea en frotis de sangre periférica o de médula ósea, aunque estos métodos no realizan la combinación de características de textura, geometría y estadística con eigenvalores de ACP como se propone en este trabajo, solo hacen la clasificación utilizando cada grupo de características. A continuación se describen brevemente algunos de estos recientes métodos.

Kan Jiang; Qing-Min Liao; Sheng-Yang Dai [36] presenta un método de segmentación de glóbulos blancos que utiliza un filtro espacio-escalar y agrupación línea divisoria de aguas (watershed). En este método, los dos componentes del glóbulo blanco (núcleo y citoplasma), son extraídos mediante métodos diferentes. Primero, se segmenta un glóbulo blanco (ROI) generando un sub-imagen, después el filtro espacio-escalar es utilizado para extraer el núcleo del glóbulo blanco. Posteriormente se agrupa mediante línea divisoria de aguas (watershed) utilizando el histograma del espacio de color HSV para extraer el citoplasma del glóbulo blanco y por último se realizan operaciones morfológicas (post-procesamiento) para obtener la conexión completa del glóbulo blanco. El objetivo de este método es extraer (del frotis de sangre periférica) un glóbulo blanco completo de un complicado fondo y segmentar cada glóbulo blanco en dos componentes morfológicas (núcleo y el citoplasma). 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.

Sánchez Segura Miriam, Rivero Jiménez René, Marsan Suárez Vianed [51] presenta un método para la clasificación de leucemia aguda linfoblástica

y mieloblástica. Se realizó mediante inmunofenotipo de células procedentes de médula ósea y de sangre periférica en 217 pacientes Del total de casos estudiados, 143 (62,44 %) fueron leucemias agudas (LA): 70 linfoides (LAL) y 38 mieloides (LAM). En 15 pacientes no se encontraron antígenos específicos de linaje y fueron clasificados como LA indiferenciadas. Se realizó la clasificación inmunológica mediante el inmunofenotipaje celular a todos los casos estudiados, con la aplicación de una batería mínima de anticuerpos monoclonales (AcMo) dirigidos tanto para antígenos mieloides como linfoides

Kyungsu, Jeon Jeonghee, Choi Wankyoo, Ho Yo-Sung [37] presentan un nuevo esquema para el análisis y clasificación automática de células en imágenes de sangre periférica. El método propuesto puede analizar y clasificar glóbulos rojos y blancos maduros (12 clases). Después de que se identifican los glóbulos rojos y blancos se extraen sus características y se clasifican mediante una red neuronal basada en el algoritmo de propagación hacia atrás (backpropagation). Se tienen quince grupos diferentes inclusive para glóbulos rojos normales y cinco categorías para glóbulos blancos (linfocito, monocito, neutrófilo, eosinófilo y basófilo). Proponen también un nuevo algoritmo de segmentación para extraer el núcleo y el citoplasma para la clasificación del glóbulo blanco, además, aplican el análisis componentes principales para reducir la dimensión de vectores de características sin afectar el desempeño de la clasificación. Obtuvieron 84% promedio de reconocimiento para 12 clases y para 4 clases 94%.

El proceso fue el siguiente:

- Para separar los glóbulos rojos, glóbulos blancos y plaquetas del fondo se aplicó un método de umbral, se clasifican por tamaño y color.
- Para separar el núcleo y citoplasma del glóbulo blanco se utiliza un híbrido basado en regiones y estados, aplicando la transformada

línea divisoria de aguas (watershed), después se combinan las regiones vecinas mediante el algoritmo k-means utilizando la información del color.

- Por último se aplica ACP y un algoritmo de difusión no linear, fusionando 3 componentes.

Saeid S. y Tracey K.M. [52], presentan un método basado en ACP y clasificación Bayesiana, para la identificación de glóbulos blancos. Para realizarlo modificaron el trabajo de Turck y Pentland en el reconocimiento de caras y lo extendieron para reconocer células. Éste utiliza un método estándar para la selección de eigenvalores Solo se usan imágenes monocromáticas. Hacen un pre-procesamiento de las imágenes por tamaño y rotación. Las eigencells son seleccionadas basadas en la minimización de similaridades. Trabajaron imágenes generadas de muestras de sangre periférica, y clasificaron todos los tipos de células generadas en el proceso de Hematopoyesis, mostrado en la figura 2.5.

Markiewicz T y Moszczyński L. [58], en su trabajo se dedica a la tarea de la generación de características para el reconocimiento automático de blastos para leucemia mielógena, mediante un proceso de segmentación, obtención de características (geométricas y estadísticas) de la ROI y su reconocimiento y clasificación mediante SVM. Trabajan con las series Eritrocitos, megacariocito, monocitos y granulocitos. Obtuvieron 48 características. Después de cada grupo de características (textura, geométricas y estadísticas) buscaron las que estaban fuertemente correlacionadas y las eliminaron. Para reducir el conjunto de características.

Theera U. N. (2005) [49], en su trabajo propone un método de segmentación dividiendo la célula en sus tres regiones, núcleo, citoplasma y fondo, este proceso lo evaluaron con las células que fueron segmentadas

manualmente por el experto. Las características que obtiene son: área, primeros y segundos momentos granulométricos del núcleo y el pico del espectro y clasifica con redes neuronales con validación cruzada (5 folders) y tiene 6 clases. Tuvieron índices de clasificación del 70.74% para su archivo de entrenamiento y 65.69% para su archivo de prueba.

El estudio del estado del arte presentado permite comprender que la segmentación de células de sangre está en continuo desarrollo.

Capítulo 4

Método Propuesto

En este capítulo se hace una descripción del método propuesto para la segmentación de glóbulos blancos, obtención de características de las ROI y clasificación de subtipos de leucemia a partir de imágenes digitales de muestras de frotis de sangre de médula ósea. Siguiendo el esquema presentado en la figura 1.2, el capítulo está organizado de la siguiente manera: En la sección 4.1 se describe la obtención de las imágenes, la sección 4.2 describe la etapa de pre-procesamiento de las imágenes. La segmentación por color se presenta en la sección 4.3. En la sección 4.4 se describe la segmentación por regiones, en la sección 4.5 se presenta la obtención de las características descriptivas de cada ROI. Y finalmente en la sección 4.6 se presenta la clasificación de las células sanguíneas de Leucemia (ROI's). En todas las secciones se ilustran los resultados generados en cada una de las etapas del método.

4.1 Obtención de Imágenes

Este trabajo inicia con la obtención de la base de datos de las imágenes digitales de subtipos de Leucemia Aguda. Las imágenes digitales de las células fueron obtenidas de la aspiración de médula ósea realizadas en el Laboratorio de Especialidades de Instituto Mexicano del Seguro Social (IMSS) en Puebla. Los frotis de médula ósea se digitalizaron utilizando un

microscopio óptico Carl Zeiss con un objetivo 100x. Este microscopio tenía una cámara digital conectada. Las imágenes obtenidas tienen una resolución 600 x 800 píxeles con 24 bits de intensidad.

La base de datos contiene 1028 imágenes de 74 pacientes, donde 415 imágenes son LAL y 613 son LAM. Estas imágenes han sido seleccionadas por el experto en el dominio.

De acuerdo con la incidencia de leucemia presentada en el IMSS de San José sólo hay 11 pacientes del subtipo L1, 8 del subtipo L2, 3 del subtipo M2, 3 del subtipo M3 y 1 del subtipo M5 (es posible obtener varias imágenes de un paciente).

En la Tabla 4 podemos observar un resumen de las regiones de interés (glóbulos blancos) de cada subtipo de leucemia con que contamos. Estas regiones de interés son nuestros ejemplos de entrenamiento y prueba

Tabla 4. Regiones de Interés para cada Subtipo de Leucemia

Leucemia Aguda	Subtipos	Imágenes	Regiones de Interés
Linfoblástica	L1	103	65
	L2	73	30
Mieloblástica	M2	57	38
	M3	48	26
	M5	11	10

Las regiones de interés fueron etiquetadas por el experto en el dominio, porque sólo el especialista conoce en que etapa de desarrollo se encuentran y cual era el tipo de célula y ya que solo se deben de tomar en cuenta mieloblastos y linfoblastos.

4.2 Pre-Procesamiento

Ya que la imagen ha sido almacenada, se inicia su pre-procesamiento, es decir las técnicas de segmentación y de reconocimiento para lograr el objetivo principal.

La segmentación de imágenes consiste en la división de una imagen en diferentes regiones de interés, cada región de interés tiene características que la distinguen de las demás y está formada por un conjunto de píxeles y éstos se agrupan con diferentes técnicas. En este trabajo utilizamos técnicas por color y detección de bordes para separar las células con leucemia (glóbulos blancos) del resto de la imagen.

La primera técnica consiste en segmentar por color la imagen utilizando la segmentación por agrupación (Clustering), ésta se realiza con el modelo de color CIELab y consiste en la extracción de los píxeles de la imagen que cumplan con las condiciones establecidas para cada color (glóbulos blancos, glóbulos rojos, plaquetas y plasma).

La segunda técnica consiste en segmentar la imagen por bordes utilizando detección de Discontinuidades (Algoritmo de Trazado de borde interno), esta segmentación se realiza sobre la imagen que anteriormente se segmentó por color y su idea principal es considerar el contexto en que se encuentran los píxeles en una vecindad local (código de cadena)

Una vez aplicadas las dos técnicas anteriores, se obtienen las regiones de interés (ROI) para posteriormente obtener las características de cada una de ellas, como se muestra en la figura 4.1.

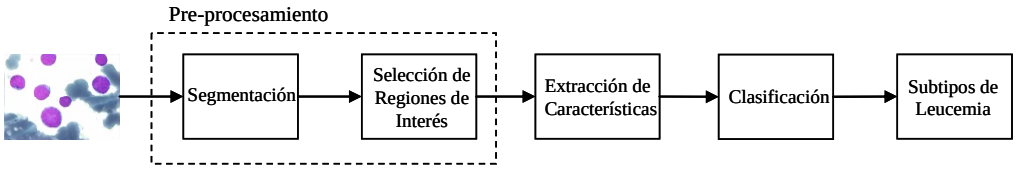


Fig. 4.1. Pre-procesamiento de Imágenes Digitales de Médula Ósea

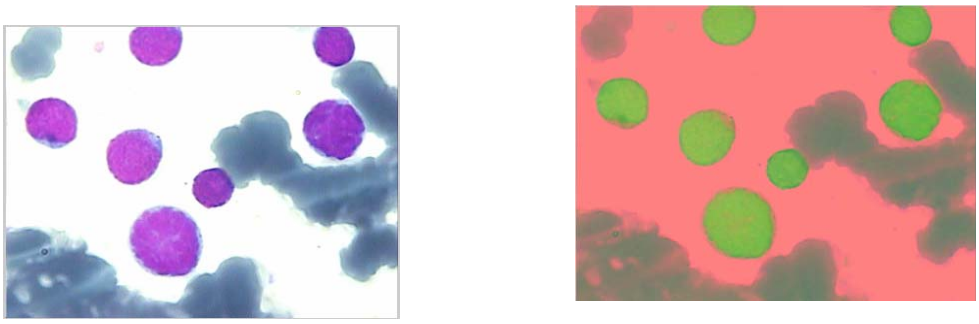
El reconocimiento y separación de regiones de interés en una imagen digital de leucemia es una tarea muy difícil porque en muchos casos las regiones de interés están juntas (se traslapan) y los bordes entre cada una de ellas son escasamente visibles; aunado a esto los colores de diferentes regiones de interés son muy similares y el citoplasma se une al núcleo o es similar al color del plasma. Esto dificulta el proceso de segmentación dado que su principal objetivo consiste en el reconocimiento automático de cada región de interés, para que posteriormente podamos obtener características que las describan.

A continuación se describe con mayor detalle el proceso de segmentación.

4.3 Segmentación por Color

Las imágenes digitales de células sanguíneas de leucemia aguda están en formato RGB, sin embargo; para el primer proceso de segmentación se convirtieron a formato CIELab; para cada imagen se genero una matriz que contiene solo 2 canales (se omitió el canal de luminosidad). Después con esta matriz se realizó la segmentación por color, con este formato se pueden percibir más fácilmente los colores de las ROI's, lo que permitió hacer una mejor agrupación por color, en cambio si hubiéramos trabajado

con el formato RGB se tendrían que trabajar tres matrices (una por canal), además como están en formato de grises, las ROI's tienden a confundirse con el fondo de la imagen y hace más difícil el proceso de segmentación. En la figura 4.2 se muestra una imagen en formato RGB y Lab. Como se puede apreciar en la figura 4.2, algunas regiones de interés (glóbulos blancos) se encuentran muy cerca de los glóbulos rojos, los bordes que las delimitan no están bien definidos y esto



a b

Fig. 4.2. Leucemia aguda linfoblástica subtipo L1. a) Formato RGB. b) Formato Lab

La figura 4.3 muestra el proceso general de segmentación por color que seguimos.

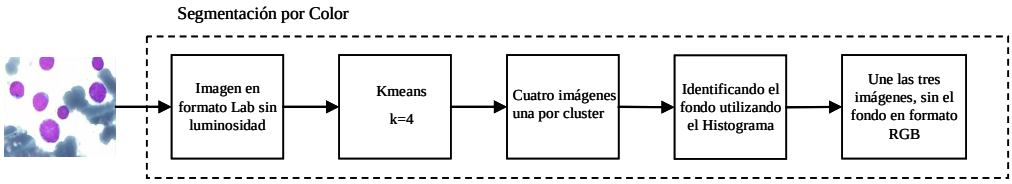
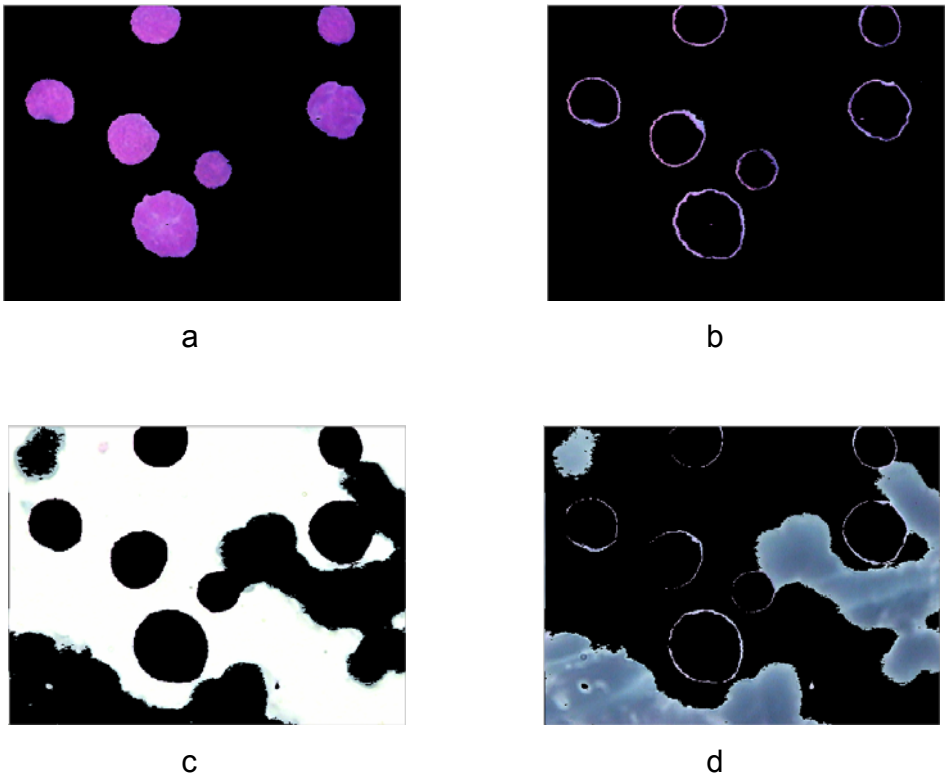


Fig. 4.3. Proceso de Segmentación por Color.

Para hacer la segmentación por color utilizamos el algoritmo K-means.

El algoritmo K-means forma grupos, los cuales son representados por k objetos. Cada uno de estos k objetos es el valor medio de los objetos que pertenecen a dicho grupo [34].

Para el algoritmo K-means [34] utilizamos como parámetros la distancia Euclidiana y k con un valor de 4, para generar 4 grupos por color. El parámetro k se ajustó a un valor de 4 porque las imágenes digitales de las células de leucemia aguda tienen 4 tonalidades principales: los glóbulos blancos, el plasma, el fondo, y los glóbulos rojos, generando cuatro imágenes, una por grupo como se muestra en la Figura 4.4 a), b), c) y d) respectivamente.



a b

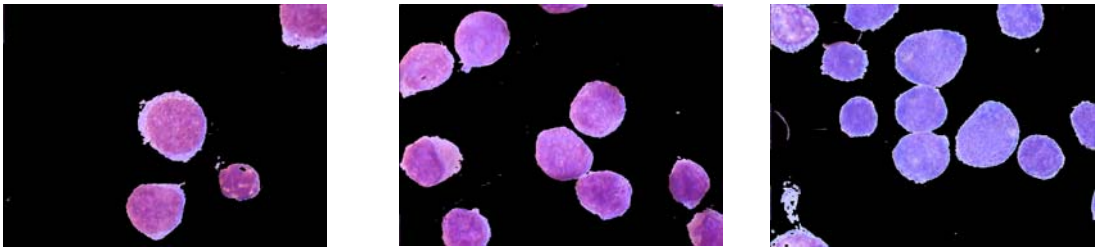
c d

Fig. 4.4. Imágenes de Leucemia aguda linfoblástica subtipo L1 separada por color, a) Glóbulos blancos. b) Plasma. c) Fondo. d) Glóbulos rojos.

Para cada una de las imágenes generadas obtenemos su histograma, que nos permitirá determinar que imágenes contienen las regiones de interés (glóbulos blancos), y cual tiene el fondo. En el último paso sólo se unirán las tres imágenes resultantes (sin incluir la imagen que tiene el fondo) en una sola en formato RGB (transformamos del formato Lab a RGB) como se muestra en la Figura 4.5. y la Figura 4.6.

a b

Fig. 4.5. Imagen de Leucemia aguda linfoblástica separada por color, sin fondo a) L1, b) L2.



a b c

Fig. 4.6. Imagen de Leucemia aguda mieloblástica separada por color, sin fondo a) M2, b) M3 y c) M5.

4.4 Segmentación de Regiones

El siguiente proceso consistió en segmentar cada región de interés (glóbulos blancos). La figura 4.7 muestra el proceso general de segmentación por región que utilizamos.

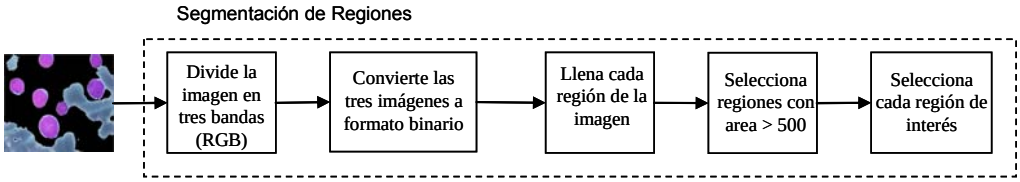


Fig. 4.7. Proceso de Segmentación de Regiones.

Para este proceso tomamos la imagen segmentada por color (sin el fondo) y la separamos en sus tres bandas (RGB). Debido a que se va a trabajar con las tres bandas, cada una de las tres imágenes generadas se convirtió a formato binario, el siguiente proceso fue detectar los bordes de cada región (detección de discontinuidades) y así numerar cada región encontrada como se muestra en la Figura 4.8 en donde las diferentes tonalidades indican las diferentes regiones encontradas. Con ayuda del experto del dominio se determinaron las regiones de interés (glóbulos blancos).

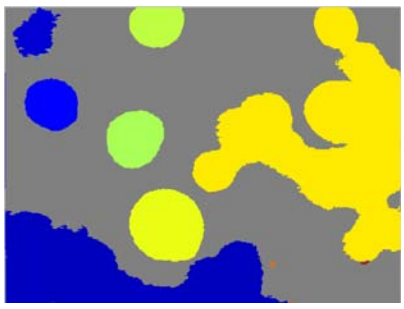


Fig. 4.8. Imagen de Leucemia aguda linfoblástica subtipo L1 segmentada por regiones.

Las regiones de interés en nuestro caso corresponden a glóbulos blancos (ver figura 4.9 y 4.10) en los que se ha identificado algún subtipo de leucemia aguda mieloblástica ó linfoblástica. El proceso anterior se repite para cada una de las imágenes de la base de datos o para cada nueva imagen de entrada a clasificar.

Una vez que hemos segmentado las regiones de interés necesitamos obtener características descriptivas de las mismas para utilizarlas como entrada a algoritmos de minería de datos y encontrar patrones que distingan entre los diferentes subtipos de leucemia. Para esto utilizamos la clasificación FAB basada en características morfológicas y histoquímicas, que toma en cuenta el nivel de maduración en el que se encuentran los blastos y la participación de las diferentes líneas celulares [30].

a b

Fig. 4.9. Región de Interés de una Imagen de Leucemia aguda linfoblástica a) L1 b) L2.

a b c

Fig. 4.10. Región de Interés de una Imagen de Leucemia aguda mieloblástica a) M2, b) M3 y c) M5.

Como se puede ver en la figura 4.10, de acuerdo a sus características morfológicas en la ROI del subtipo 1 casi no se perciben los nucleolos, mientras en la ROI del subtipo 2 se perciben nucléolos. En la ROI del subtipo M2 hay más citoplasma mientras que las ROI de los subtipos M3 y M5 tienen escaso citoplasma. En el subtipo M5 es un poco más grande la ROI.

Una vez localizada en cada imagen su ROI, se tienen que obtener características que discriminen cada una de estas ROI, como se presenta en la siguiente sección.

4.5 Obtención de Características Descriptivas de ROI's

Para poder clasificar las diferentes regiones de interés (glóbulos blancos en nuestro caso) es importante obtener características descriptivas que ayuden a discriminar los diferentes subtipos de ellas. En este trabajo utilizamos tres tipos de características de textura, geométricas y estadísticas.

Una característica de una imagen es un atributo primitivo distintivo de la misma. Algunas características están definidas por la apariencia visual de una imagen, mientras que otras resultan de alguna manipulación específica.

Las características de la imagen son importantes en el aislamiento de regiones que tienen propiedades comunes dentro de una imagen (segmentación de imágenes) y la posterior identificación o etiquetado de tales regiones (clasificación de imágenes).

Para obtener las características convertimos la imagen RGB que contiene la región de interés a tonos de grises, de esta nueva imagen obtuvimos 30 características como el Perímetro, Área Eje Mayor, Eje Menor, Umbral de Gris, Solidez, Total de Píxeles, la Media y la Varianza entre otras.[21, 55] (Ver Tabla 5).

Algunas de éstas fueron obtenidas con procedimientos de la librería para imágenes de Matlab [42] (regionprops.- calcula un conjunto de propiedades para cada región de interés etiquetada), y otras más se programaron con comandos (también en Matlab [42]). En imágenes digitales las características de textura se representan por la interrelación entre los arreglos de los píxeles de la imagen, y se ven como cambios en la intensidad o los tonos de grises. Las características estadísticas se generan sobre la base de la distribución de la imagen. Mucha información de los glóbulos blancos está contenida en la forma geométrica de la célula. Aplicando estos métodos de generación de características, se obtuvieron 30 para cada región de interés.

El propósito de este trabajo es encontrar características que ayuden a discriminar entre las diferentes clases para lograr una buena clasificación. En busca de otras características que nos ayuden a mejorar la precisión de la clasificación utilizamos la técnica de análisis de componentes principales para encontrar los eigenvalores que describan mejor a una región de interés como se muestra en la siguiente sección (4.5.1)

Tabla 5. Características obtenidas de cada Regiones de Interés

Características	Descripción	Tipo
1.- Perímetro	Es la longitud de su frontera.	Geométrica
2.- Distancia	Es la distancia de una línea recta entre dos píxeles (distancia Euclidiana)	Geométrica
3.- Área	Es el número de píxeles contenidos dentro de su frontera.	Geométrica
4 y 5 .-Centroide	Es el centro de la región de interés (coordenada x y coordenada y).	Geométrica
6, 7, 8 y 9.- Esquinas	Son las cuatro esquinas del rectángulo que contiene a la región de interés	Geométrica
10.- Eje mayor	Es la longitud del eje mayor de la elipse que atraviesa a la región de interés	Geométrica
11.-Eje menor	Es la longitud del eje menor de la elipse que atraviesa a la región de interés	Geométrica
12.- Excentricidad	Número escalar que da la excentricidad la elipse que atraviesa a la región de interés	Geométrica
13.-Orientación	Es el ángulo que existe entre el eje mayor y el eje menor de la elipse que atraviesa a la región de interés	Geométrica
14.- ConvexArea	Es el número de píxeles en la región de interés	Geométrica
15.- EulerNumber	Es un número escalar que es igual 1 menos el número de agujeros de la región de interés	Geométrica
16.- EquivDiameter	Es el diámetro del circulo con la misma área de la región de interés	Geométrica
17.- Solidez	Es la proporción de píxeles del Área/ConvexArea	Geométrica
18.- Extensión	Es la proporción de píxeles del Área/Perímetro	Geométrica
19.- PoliEjeX	Es el total del número de píxeles de la coordenada x del polígono que contiene a la región de interés	Geométrica
20.- PoliEjeX	Es el total del número de píxeles de la coordenada y del polígono que contiene a la región de interés	Geométrica
21 y 22 .- Entropía	Mide la uniformidad del histograma	Textura
23.- Compactes	Es la proporción del Perímetro^2 /(4*Área*pi)	Geométrica
24.- Umbral de Gris	Es el umbral de la región de interés	Textura
25.- Histograma	Es la suma total del histograma de la región de interés	Textura
26.-Maxh	Es el valor máximo del histograma de la región de interés	Textura
27.-Minh	Es el valor mínimo del histograma de la región de interés	Textura
28.- Media	Es valor medio de los niveles de gris de la región de interés. Indica el brillo o luminosidad de la región de interés	Estadística
29.-Desviación Estándar	Es la media del contraste, es decir la variación de la información en la región de interés	Estadística
30.- Varianza	Es la varianza de los niveles de gris de la región de interés	Estadística

4.5.1 Obtención de Componentes Principales

El siguiente proceso es obtener otras características de tipo estadístico que en nuestro caso son los valores propios (eigenvalores) del Análisis de Componentes Principales.

El conjunto de valores propios se tomaron como características descriptivas de las RIO's. Y solo seleccionamos los valores propios que acumulen el 80% de variabilidad, que para nuestro caso fueron 10 valores propios de cada banda (30 eigenvalores en total por cada región de interés) para cumplir con esta condición.

Para poder obtener los eigenvalores, primero tomamos la ROI de la imagen, como esta en formato RGB, dividimos en cada una de las bandas de la imagen, después se hace el calculo de la matriz de correlación de la ROI (para cada banda) y por ultimo se calculan los valores propios (eigenvalores) y vectores propios (eingenvectors) de cada banda resolviendo la ecuación característica (formula 10). Todo los anterior se realizó con comandos de Matlab [42] y utilizando la metodología explicada anteriormente para Análisis de Componentes Principales.

Una vez que obtuvimos los eigenvalores hicimos experimentos con las diferentes características obtenidas de las regiones de interés (geométricas, textura, estadísticas, y eigenvalores) con el objetivo de obtener la mejor precisión posible al clasificar los diferentes subtipos de leucemia aguda linfoblástica y mieloblástica.

4.6 Clasificación

El reconocimiento de patrones tiene como objetivo la clasificación de objetos (ROI's) en un cierto número de clases con base en un conocimiento a priori o información extraída de los patrones. Los patrones a clasificar suelen ser grupos de medidas (atributos) u observaciones.

La clasificación supervisada usa un conjunto de aprendizaje, del cual ya se conoce la clasificación de la información a priori y se utiliza para entrenar al sistema.

En identificación de las ROI's, se extraen ciertas características de cada imagen antes de ser clasificadas. Se busca seleccionar aquellas características relevantes para poder discriminar mejor, de modo de no trabajar con la imagen original. Después la clasificación se realizará mediante el análisis de las características de obtenidas de las ROI's.

En este trabajo se usaron diferentes clasificadores para abordar el reconocimiento de las células sanguíneas de Leucemia y nos enfocamos a los clasificadores con que cuenta la herramienta WEKA: BayesNet, NaiveBayes, NaiveBayesSimple, Logistic, MultilayerPerceptron, SimpleLogistic, SMO, IB1, IBk, KStar, LWL, ADTree, J48, LMT, RandomForest, RandomTree, JRip, OneR,Part y como ensamble AdaBoostM1. También se aplico selección de atributos con ChiSquaredAttributeEval, GainRatioAttributeEval, InfoGainAttributeEval, OneRAttributeEval

El objetivo de la selección de atributos es identificar, mediante un conjunto de datos que poseen unos ciertos atributos, aquellos atributos

que tienen más peso a la hora de determinar si los datos son de una clase u otra. Dependiendo del método seleccionado, será el encargado de evaluar cada uno de los atributos y dotarlos de un peso específico.

Los siguientes algoritmos de selección de atributos son evaluadores de atributos individuales y a cada uno se le aplicó también el método Ranker, que devuelve una lista ordenada de los atributos según su calidad:

- **ChiSquaredAttributeEval**: calcula el valor estadístico Chi-cuadrado de cada atributo con respecto a la clase y así obtiene el nivel de correlación entre la clase y cada atributo.
- **GainRatioAttributeEval**: evalúa cada atributo midiendo su razón de beneficio con respecto a la clase.
- **InfoGainAttributeEval**: evalúa los atributos midiendo la ganancia de información de cada uno con respecto a la clase. Antes discretiza los atributos numéricos.(14)
- **OneRAttributeEval**: evalúa la calidad de cada atributo utilizando el clasificador OneR, el cual usa el atributo de mínimo error para predecir, discretizando los atributos numéricos.

El algoritmo de aprendizaje fue aplicado a los datos antes y después de la selección de los atributos, y de esta forma comparar el porcentaje de células sanguíneas de Leucemia mal clasificadas considerando todas las variables, con cada uno de los porcentajes aportados por la clasificación luego de la selección de los atributos. La prueba se realizó con todos los clasificadores antes mencionados, pero como se pudo constatar al hacer selección de atributos no mejoró la clasificación, por lo que se opto por tomar todos los atributos (30) pero

con el método Ranker (se tomo el umbral que tiene por defecto -1.7976931348623157E308) para que los ordenara según su calidad.

Como se mencionó anteriormente tenemos clases desbalanceadas, por lo cual se tuvo que aplicar el proceso SMOTE para poder balancear las clases minoritarias, antes de iniciar la clasificación.

El proceso de SMOTE [15, 53] replica ejemplos de la clase minoritaria. Algoritmo: para cada ejemplo de clase minoritaria, introduce ejemplos sintéticos a lo largo de los segmentos que unen a cualquiera(o todos) los vecinos cercanos de la clase minoritaria k, como se muestra en la Figura 4.11.

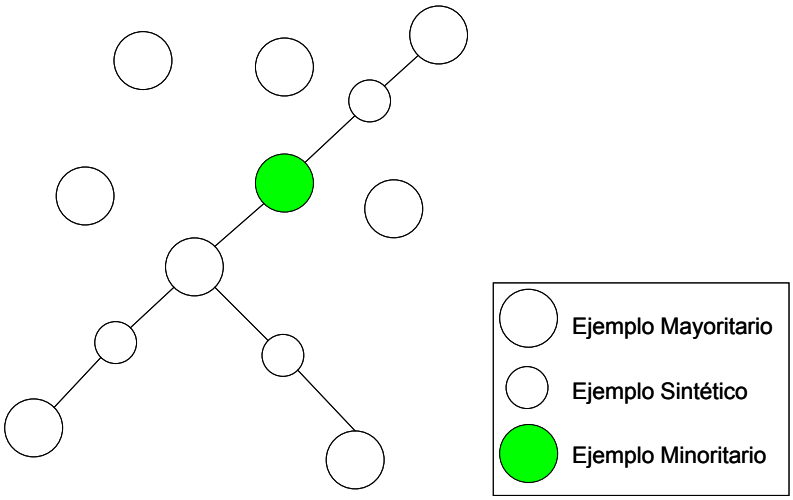


Fig. 4.11. Algoritmo SMOTE.

El proceso para generar ejemplos sintéticos (SMOTE) se realizó de la siguiente manera:

- Primero se aparta el 10% de los ejemplos de forma aleatoria de cada unos de los subtipos L_2, M_3 y M_5, estos servirán para probar el modelo generado.

- Del 90% restante de los ejemplos de cada uno de los subtipos antes mencionados se dividen en 9 grupos y a cada grupo se le aplica el algoritmo de SMOTE.
- Por último se unen para cada subtipo el 90% de los ejemplos originales con sus correspondientes ejemplos generados por el algoritmo de SMOTE. Para generar así generar los archivos de entrenamiento que van a utilizar para crear los modelos.

Del proceso anterior obtuvimos los siguientes porcentajes:

- Para L_2 se le aplicó el 100% de SMOTE y se generaron 60 ejemplos, para M_3 se le aplicó el 100% de SMOTE y se generaron 52 ejemplos y para M_5 se le aplicó el 400% de SMOTE y se generaron 40 ejemplos

A los subtipos L_1 y M_2 no se les aplicó SMOTE, se tomaron como base de clase mayoritaria para poder balancear el resto de los subtipos con los de la clase mayoritaria.

Se realizaron diferentes experimentos con clases balanceadas y desbalanceadas:

- Por tipos de leucemia (LAL, LAM)
- Todos los subtipos de leucemia juntos (L_1, L_2, M_2, M_3, M_5)
- Por grupos de subtipos de Leucemia (L_1,L_2 y M_2, M_3, M_5)
- En cascada, primero por tipos de leucemia (LAL, LAM) y luego por grupos de subtipos de Leucemia (L_1, L_2 y M_2, M_3, M_5)
- En cascada, primero por tipos de leucemia (LAL, LAM) y luego por grupos de subtipos de Leucemia con clases binarias (L_1, L_2 y M_2, M's y M_3, M's y M_5 M's)

Pero solo se presentan los mejores resultados obtenidos al clasificar en la siguiente sección.

Capítulo 5

Experimentos

5.1 Tipos de Experimentos

En esta sección describimos los experimentos que se realizaron de acuerdo a las características obtenidas, utilizando diferentes algoritmos de aprendizaje implementados en la herramienta Weka [63] para la clasificación. Es importante mencionar que la base de datos presenta el problema de clases desbalanceadas entre subtipos de leucemia y por tanto en nuestros experimentos incluimos técnicas de sobre-muestreo [15, 53].

Hicimos tres tipos de experimentos para la clasificación de tipos y subtipos de leucemia aguda, en el primer experimento solo tomamos en cuenta las características de textura, geométricas y estadísticas (30 en total). En este experimento se utilizó una selección de atributos para trabajar con menos de las 30 características que obtuvimos, pero no hubo mejores resultados y por eso mostramos aquí el experimento con la totalidad de características; el segundo experimento solo utilizamos características generadas por ACP (eigenvalores) que también son 30 y por último en el tercer experimento considera todas las características: eigenvalores, textura, geométricas y estadísticas.

Estos tres experimentos se realizaron para clasificar tipos y subtipos de Leucemia. La evaluación de los resultados se hizo utilizando la técnica de validación cruzada con significancia estadística.

5.2 Clasificación basada en Características Geométricas, de Textura y Estadísticas

Para el primer experimento, antes de iniciar la clasificación se le aplicó al archivo de características un método de filtro supervisado de selección de atributos, con un evaluador de ganancia de información y búsqueda Ranker, donde los atributos se ordenan de acuerdo a su correlación con la clase y ese archivo generado fue el que se ocupó para la clasificación.

En nuestra clasificación de subtipos de leucemia contamos con 5 clases, 2 que corresponden a la familia linfoblástica (L1 y L2) y tres de la familia mieloblástica (M2, M3, y M5). Es importante señalar que la clasificación de subtipos se hizo por familia, es decir; primero hacemos la clasificación por familias para separar los ejemplos de leucemia linfoblástica de los ejemplos de leucemia mieloblástica y posteriormente hacemos una clasificación por subtipos utilizando ejemplos de una sola familia. Por tanto, creamos dos grupos de datos, el primer grupo para los subtipos L1, y L2 y el segundo para los subtipos M2, M3 y M5, porque el objetivo del trabajo propuesto fue de clasificar subtipos de leucemia por lo tanto se tuvo que clasificar primero por familia.

La clasificación por familias de Leucemia fue: 82.3% de precisión para Linfoblástica (L) y 91.7% de precisión para Mieloblástica (M) como se muestra en la tabla 6.

Tabla 6. Resultados de clasificación de características de textura, geométricas y estadísticas para tipos de leucemia

Tipos de Leucemia Aguda	Clasificador	Precisión
L's, M's	lazy.IBk -K 4	82.35%

Para realizar la prueba de los subtipos M1, M2 y M3 se hizo la clasificación con clases binarias es decir se tomo un subtipo como una clase y los otros dos subtipos como la clase restante, porque daban mejores resultados que si se trabajan las 3 clases de la familia mieloblástica, porque los atributos que se calcularon a nivel morfológico son parecidos entre si. Con respecto a los resultados obtenidos, se concluye que se lograron clasificar mejor los subtipos de la leucemia mieloide, esto puede ser consecuencia de que las atributos en los subtipos de leucemia linfoblástica son muy similares entre si, además los ensambles arrojan mejores porcentajes de clasificación.

Tabla 7. Resultados de clasificación de características de textura, geométricas y estadísticas para subtipos de Leucemia

Subtipos de Leucemia Aguda	Clasificador	Precisión
L1, L2	RandomCommittee.RandomForest	86.40%
M2 y resto de Ms	AdaBoostM.IBk	91.53%
M3 y resto de Ms	IBk	88.46%
M5 y resto de Ms	AdaBoostM.BayesNet.TAN	95.38%

5.3 Clasificación basada en Eigenvalores (ACP's)

Para esta prueba cabe mencionar que los atributos (eigenvalores) ya están ordenados de acuerdo a la variabilidad que cada uno representa, por lo cual no fue necesario aplicar selección de atributos.

En la tabla de resultados 8 obtenemos nuevamente una mejora en la clasificación de los tipos de la leucemia Mieloide, pero de nuevo las características de los tipos de Leucemia linfoblástica son muy similares y es más difícil de discriminar. Los porcentajes de clasificación son similares a los de la clasificación de las características de textura, geométricas y estadísticas, aún sin haber realizado el balance de las clases.

En la clasificación de los tipos de Leucemia Mieloide los clasificadores de ensambles obtuvieron mejores resultados.

El promedio de la clasificación de entre L1 y L2 es 87.2% y de M2, M3 y M5 es 93.40%.

Tabla 8. Resultados de clasificación de eigenvalores para tipos de leucemia

Tipos de Leucemia Aguda	Clasificador	Precisión
L's y M's	RandomForest	87.451%
	KStar	86.274%

Tabla 9. Resultados de clasificación de eigenvalores para subtipos de leucemia

Subtipos de Leucemia Aguda	Clasificador	Precisión
L1, L2	AdaBoost.IB1	90.76%
M2 y resto de Ms	AdaBoost.KStar	90.00%
M3 y resto de Ms	AdaBoost.SMO	98.46%
M5 y resto de Ms	IBk -K 2	97.69%

5.4 Clasificación basada en Características Geométricas, de Textura, Estadísticas y Eigenvalores

En este último experimento se unieron todas las características, los eigenvalores con las características de textura, geométricas y estadísticas y en total obtuvimos 60 atributos, 30 atributos de eigenvalores y 30 de características de textura, geométricas y estadísticas.

Para este experimento los mejores resultados los obtuvimos con los algoritmos: LMT, SMO, y el ensamble AdaBoost.SMO. Aunque se hicieron pruebas con otros algoritmos pero solo presentamos los mejores resultados en la tabla 10.

Como se puede ver en la tabla 11 se obtuvo mejor precisión para los subtipos de leucemia aguda mieloblástica. Y cuando utilizamos los 60 atributos (eigenvalores con textura, geométricos y estadísticos), obtenemos mejores resultados en L1, L2, M2 y M5 (la mayoría de las clases).

El promedio de la clasificación de entre L1 y L2 es 89.6% y de M2, M3 y M5 es 93.33%.

Tabla 10. Resultados de clasificación de características de textura, geométricas y estadísticas y eigenvalores para tipos de leucemia

Tipos de Leucemia Aguda	Clasificador	Precisión
L's y M's	RandomForest	88.2353%

Tabla 11. Resultados de clasificación de características de textura, geométricas y estadísticas y eigenvalores para subtipos de leucemia

Subtipos de Leucemia Aguda	Clasificador	Precisión
L1, L2	trees.LMT	89.6%
M2 y resto de Ms	functions.SMO	92.30%
M3 y resto de Ms	AdaBoost.IB1	88.46%
M5 y resto de Ms	functions.SMO	99.23%

Capítulo 6

Conclusiones

6.1 Conclusiones

Con base a los experimentos realizados y los resultados obtenidos en el presente trabajo se pueden hacer las siguientes conclusiones.

Como primera conclusión con respecto al proceso de clasificación de subtipos de leucemia aguda comprobamos que el uso de características descriptivas (geométricas, textura, y estadísticas) por un lado y eigenvalores (ACP) por el otro produce resultados con buena precisión; alrededor de 85% para los subtipos L1 y L2 y alrededor del 91% para los subtipos M2, M3 y M5; aunque siempre se obtuvieron mejores resultados utilizando eigenvalores (ACP), a excepción de la clasificación del subtipo M3. Según los expertos en el dominio de leucemia esta precisión es suficiente para un sistema de apoyo al diagnóstico médico.

Para todos los subtipos de leucemia aguda a clasificar, los mejores porcentajes de clasificación fueron de leucemia mieloide.

Al comparar los resultados obtenidos para la clasificación por familias de leucemia con los resultados obtenidos por otros métodos descritos en el estado del arte, éstos son muy similares a los reportados por nosotros, aunque realiza tareas diferentes al analizar diferentes líneas celulares (por tanto no son directamente comparables). Cabe mencionar que no

encontramos ningún trabajo que realice clasificación de subtipos de leucemia como lo hacemos nosotros.

Otro experimento que no se había hecho es el uso de ambos grupos de características (las características descriptivas con los eigenvalores para describir regiones de interés) con lo que mejoramos la precisión en la clasificación comparado con el uso de los mismos tipos de características por separado. Este resultado es importante porque se puede utilizar no sólo con la base de datos de leucemia sino en general para cualquier tipo de regiones de interés. La validación de estos resultados se hizo con pruebas de hipótesis en donde se verificó que las diferencias en precisión son estadísticamente significativas. Por último, también se validó la clasificación con resultados de pruebas de citometría de flujo que nos proporcionó el experto en el dominio.

Los resultados anteriores nos llevan a la conclusión de que la visión por computadora junto con el aprendizaje automático (minería de datos) se pueden utilizar para ayudar al médico en tareas de apoyo al diagnóstico médico de subtipos de leucemia aguda linfoblástica y mieloblástica. Este tipo de apoyos también serán de utilidad para entrenar a estudiantes de medicina en la práctica de detección de leucemia.

6.2 Trabajos Futuros

Existe mucho trabajo por realizar para aplicar los resultados obtenidos hasta el momento.

En primer lugar es necesario realizar más experimentos con técnicas de sobremuestreo, no sólo para las clases minoritarias porque también

contamos con pocos ejemplos de las clases mayoritarias. También es necesario obtener ejemplos de subtipos que no se contemplaron en este trabajo y más ejemplos de los que si se tomaron en cuenta. Por otro lado queremos encontrar otros tipos de características que nos ayuden a mejorar la precisión en la clasificación hasta ahora alcanzada. Por ejemplo, podemos encontrar algunas características espaciales que podrían ayudar, por ejemplo; el citoplasma de la célula en algunos casos rodea a todo el núcleo (relación contiene) y en otros casos sólo es adyacente a una parte del mismo. Por último es necesario trabajar más en la parte de segmentación de las regiones de interés para hacer posible la obtención de las características espaciales y mejorar las que ya hemos obtenido.

Otro trabajo futuro se encuentra el integrar el método propuesto en un sistema de diagnóstico para que sirva de apoyo al médico en el dictamen y prescripción del tratamiento o para crear un sistema de entrenamiento para hematólogos. Esto traerá como beneficio la detección de leucemia en una etapa temprana para determinar el tipo de tratamiento que se debe asignar al paciente.

Es necesario también realizar el proceso con sangre periférica (en lugar de médula ósea), ya que con solo una simple muestra de sangre y un análisis morfológico de la misma, este sistema se podría ocupar en lugares donde no hay los instrumentos necesarios para realizar un estudio más complejo como es la citometría de flujo.

Glosario

Acido desoxirribonucleico (ADN): ácido nucleico constituido por un gran número de nucleótidos unidos y dispuestos en dos hélices. Constituye el material cromosómico y contiene toda la información hereditaria correspondiente a la especie.

Anticuerpos Monoclonales: son sustancias producidas en laboratorios, que reconocen y se unen a glóbulos blancos específicos (tal como una proteína) existentes en la superficie de una célula cancerígena.

Basofilia: se define como el aumento en el número absoluto de basófilos circulantes por encima 200 células/mm3.

Eritropoyesis: Es el proceso de formación de los eritrocitos. A partir de las células madre o pluripotentes, con capacidad para producir todas las células de la sangre.

Frotis de sangre periférica: es la extensión de sangre sobre una lámina de vidrio, se tiñe con el colorante May-Grünwald y permite el estudio de la morfología de los elementos de la sangre.

Frotis: es el análisis exhaustivo de la sangre humana o de cualquier otro animal empleando tintes y recuento al microscopio.

Hueso esternón (Sternum): es un hueso del tórax, plano, impar, central y simétrico, compuesto por varias piezas soldadas (esternebras).

Inmunofenotipo: Los diferentes tipos de leucemia tienen una proteína y/ o carbohidrato único llamado antígeno que se encuentra en la superficie o dentro de la célula. Ciertos antígenos están correlacionados con unas características específicas de la enfermedad, lo cual ayuda a clasificar la leucemia y a definir la opción óptima de tratamiento. La detección del antígeno específico se llama determinación del fenotipo inmune. La prueba de laboratorio llamada inmunohistocompatibilidad (IHC) es capaz de verificar una multitud de antígenos en una muestra de sangre o tejido.

Leucemia linfoblástica aguda (LAL): es una enfermedad maligna o cáncer de la sangre que se caracteriza por el crecimiento rápido e incontrolado de glóbulos blancos inmaduros anormales conocidos como linfoblastos.

Leucemia mielógena aguda (LAM): es un cáncer de la médula ósea y la sangre que se caracteriza por el crecimiento incontrolado de los glóbulos blancos inmaduros conocidos como mielocitos.

Linfoblasto: es una célula que semeja un mieloblasto en cuanto a su estructura general, mide unos 15 a 20 µm de diámetro. Posee un citoplasma no granular que se tiñe de azul oscuro en la periferia y más claro en el centro. El núcleo es grande, pues suele ocupar las cuatro quintas partes del área celular, y la cromatina dispuesta en forma reticular tiende a ser punteada. Por lo general sólo contiene uno o dos nucléolos.

Linfocitos B: son glóbulos blancos que generan anticuerpos para luchar contra bacterias y toxinas. Reconocen a las bacterias y se incorporan a éstas. Permitiendo que los glóbulos blancos: granulocitos y los monocitos puedan destruir a las bacterias cuando los linfocitos están unidos a éstas.

Linfocitos T: son glóbulos blancos que regulan procesos inmunológicos de rechazo a infecciones virales, cáncer y transplantes, reconocen a las células infectadas por los virus y las destruyen con ayuda de los macrófagos.

Mieloblasto: es una célula y su tamaño está entre 15 y 20 µm. El citoplasma puede ser agranular o exhibir unos pocos gránulos azurófilos según la etapa de desarrollo. Es moderadamente azul intenso en una región tintorial que puede ser despareja, que a menudo es más clara que la región perinuclear. El núcleo es redondo u oval y ocupa cerca de las cuatro quintas partes del área total de la célula. Puede haber hasta 6 nucléolos, pero lo usual son de dos a cinco; esto nucléolos son de tamaño mediano y suelen estar muy bien definidos, con un borde de cromatina bien marcado.

Modelo HSV: del inglés Hue, Saturation, Value – Tonalidad, Saturación, Valor, también llamado HSB (Hue, Saturation, Brightness – Tonalidad, Saturación, Brillo), define un modelo de color en términos de sus componentes constituyentes en coordenadas cilíndricas.

Neoplasia (*nuevo crecimiento* en griego) es el proceso de proliferación anormal de células en un tejido u órgano que desemboca en la formación de un tumor (neoplasma).

Nucleolo: es una estructura densa, aparecen a razón de dos o tres por célula, aunque eso dependerá del tipo celular y de la actividad de ésta. Morfológicamente, el nucleolo suele ser esférico pero puede adoptar formas muy irregulares. Suelen encontrarse en el centro del núcleo o ligeramente desplazados hacia la periferia. Su tamaño puede ser también muy variable pero suele oscilar entre una y dos micras.

Organelas (del griego organon = herramienta): Estructuras subcelulares que realizan determinadas funciones (generalmente están rodeadas por membranas y se las encuentra en las células eucariotas) p.ej.: mitocondrias, cloroplastos.

Tomografía computarizada (TC) [9]: Es un método imagenológico que usa rayos X para crear imágenes transversales detalladas del cuerpo y se utiliza para estudiar los vasos sanguíneos, identificar masas y tumores, incluyendo cáncer o guiar a un cirujano hacia el área correcta durante una biopsia

Referencias

1. Araujo Basilio, *Aprendizaje Automático: conceptos básicos y avanzados*, Prentice Hall, 1st ed, 23, 25, 41, 42, 59, 77, 101, 133, 163, 261, 295, 379

2. Aprendizaje no supervisado y análisis de agrupamientos, http://iie.fing.edu.uy/ense/asign/recpat/material/agrupamiento-pres.pdf Fecha de Consulta: 8 de Febrero de 2007

3. Barrera RLM, Drago SME, Pérez RJ, Zamora AC et al, "Citometría de flujo: vínculo entre la investigación básica y la aplicación clínica. Revista del Instituto Nacional de Enfermedades Respiratorias" Mex 2004; 17 (1): 42-55

4. Behrman RE. Nelson Textbook of Pediatrics. 17th ed. Philadelphia, Pa: WB Saunders; 2004; 1695-1697.

5. Bennett J. M. et al. *Proposals for the classification of the acute leukemias. French-American-Brithis (FAB) co-operative group*, Br J Haematol.; Vol 4, No. 33(1976), 451-458

6. Bennett JM, Catovsky D, Daniel MT, Flandrin G, Galton DA, Gralnick HR, et al. "Proposal for the recognition of minimally differentiated acute myeloid leukemia (AML-M0)". Br J Haematol 1991; 78: 325-329. [Medline]

7. Bennett JM, Catovsky D, Daniel MT, Flandrin G, Galton DA, Gralnick HR, et al. "Proposed revised criteria for the classification of acute myeloid leukemia." A report of the French- American- British Cooperative Group. Ann Intern Med 1985; 103: 620-625. [Medline]

8. Biblioteca de Salud GEX, http://stanthony.cht-info.com/HealtheraApps/Healthlibrary/OneLibrary.aspx?id=622&selLangs=ECS&lang=SP Fecha de Consulta: 9 de Febrero de 2007

9. Cancer Consultants,
http://patient.spanish.cancerconsultants.com/CancerTreatment_Leucemia.aspx?LinkId=56053
Fecha de Consulta: 8 de Febrero de 2007

10. Cáncer,
http://www.elmundo.es/elmundosalud/especiales/cancer/leuc_agudas3.html
Fecha de Consulta: 13 de Febrero de 2007

11. Cáncer,
http://www.cancerinfo.es/index.php?textoid=21&orden=4
Fecha de Consulta: 13 de Febrero de 2007

12. Cáncer,
http://www.elmundo.es/elmundosalud/especiales/cancer/leuc_agudas4.html
Fecha de Consulta: 6 de Febrero de 2007

13. Cáncer: cifra record,
http://esp.mexico.com/lapalabra/una.php?idarticulo=8871
Fecha de Consulta: 28 de Febrero de 2007

14. Citometría de flujo Fundamentos y Aplicaciones,
http://www.citometriadeflujo.com/HTML/fundamentos%20frame.htm
Fecha de Consulta: 12 de Marzo de 2007

15. Clasificación con Datos no Balanceados,
http://sci2s.ugr.es/docencia/doctoM6/M6-Clasificacion%20con%20datos%20no%20balanceados.pdf
Fecha de Consulta: 23 de Marzo de 2007

16. Descripción General de la Sangre,
http://www.healthsystem.virginia.edu/UVAHealth/peds_hrnewborn_sp/ovrblood.cfm
Fecha de Consulta: 9 de Febrero de 2007

17. Digital Image Process,
http://iria.pku.edu.cn/~jiangm/courses/dip/html/node127.html
Fecha de Consulta: 7 de Febrero de 2007

18. Duda Richard, Hart Peter, *Pattern Classification and Scene Analysis*, A Wiley-Interscience Publication, 1st ed, 31, 44, 45, 57, 85, 95, 103, 130, 141, 327, 341, 352

19. Duda Richard, Hart Peter, Stork David, *Pattern Classification*, A Wiley-Interscience Publication, 2nd ed, 20, 64,84, 161, 215, 282, 350, 394, 411, 431

20. Estadística INEN, Abril,
http://www.inen.sld.pe/intranet/estadepidemiologicos.htm
Fecha de Consulta: 20 de Abril de 2007

21. Fotogrametría Digital,
http://books.google.com/books?id=k561-RT8Ew4C&pg=PA66&lpg=PA66&dq=imagen+desviacion+estandar&source=web&ots=QrbV6ZgMLM&sig=f8E-bcMP5JI3k8-LM58pnds4KVY#PPA67,M1
Fecha de Consulta: 16 de Marzo de 2007

22. González R., Woods R., Eddins S., *Digital Image Processing Using MATLAB*, Pearson-Prentice Hall,1st ed, 2, 3, 494, 194, 195, 206, 463, 478, 485

23 González R., Woods R., Eddins S., *Digital Image Processing*, Pearson-Prentice Hall,1st ed, 15, 34, 148, 282, 519, 567, 643, 693

24. Harris NI, Jaffe ES, Diebold J, Flandrin G, Müller- Hermelink HK, Vardiman J, et al. "World Health Organization classification of neoplastic diseases of the hematopoietic and lymphoid tissues: report of the Clinical Advisory Committee meeting"- Airlie House, Virginia, November 1997. J Clin Oncol 1999; 17: 3.835-3.849.

25. Hematopoietic stem cell,
http://www.answers.com/topic/hematopoietic-stem-cell
Fecha de Consulta: 29 de Marzo de 2007

26. Hoffman R, Benz EJ, Shattil SS, et al." Hematology: Basic Principles and Practice". 4th ed. Orlando, Fl: Churchill Livingstone; 2005:2656-2657.

27. Imagen,
http://www.healthsystem.virginia.edu/uvahealth/peds_genetics_sp/images/ss_0090.gif
Fecha de Consulta: 1 de Febrero de 2007

28. Imagen,
http://recursos.cnice.mec.es/biosfera/alumno/3ESO/aparato_circulatorio/Dibujos/Circul5-1-1.jpg
Fecha de Consulta: 21 de Febrero de 2007

29. Imagen,
http://recursos.cnice.mec.es/biosfera/alumno/2bachillerato/inmune/imagenes/organos/medula3.jpg
Fecha de Consulta: 21 de Febrero de 2007

30. Instituto Nacional de Estadística Geográfica e Informática,
http://www.inegi.gob.mx/inegi/default.aspx
Fecha de Consulta: 5 de Febrero de 2007

31. Instituto Nacional del Cáncer,
http://www.cancer.gov/Templates/db_alpha.aspx?CdrID=45107&lang=spanish
Fecha de Consulta: 6 de Marzo de 2007

32. Internacional Agency for Research on Cancer,
http://www-dep.iarc.fr/
Fecha de Consulta: 7 de Febrero de 2007

33. Introducción al Procesamiento Digital de Imágenes,
http://delta.cs.cinvestav.mx/~fraga/Cursos/PDI/cap1.pdf
Fecha de Consulta: 9 de Febrero de 2007

34. J. B. MacQueen (1967): "Some Methods for classification and Analysis of Multivariate Observations, Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability", Berkeley, University of California Press, 1:281-297

35. K. Pratt, William. *Digital Image Processing*. John Wiley & Sons, 1991, 1st ed, 262

36. 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

37. Kim Kyungsu, Jeon Jeonghee, Choi Wankyoo, Ho Yo-Sung, **"**Automatic Cell Classification in Human's Peripheral Blood Images Based on Morphological Image Processing." 14th Australian Joint Conference on Artificial Intelligence: Advances in Artificial Intelligence, 2001 p. 225 – 236.

38. La Hematología y las Enfermedades de la Sangre,
http://www.healthsystem.virginia.edu/UVAHealth/adult_blood_sp/blood.cfm
Fecha de Consulta: 9 de Febrero de 2007

39. Leucemias Agudas,
http://mipagina.aol.com.mx/__121b_0PQim+vkRtjkZ1gR2T9Lkr5ayQk+Nzh7S5lKlSxW25c=
Fecha de Consulta: 15 de Enero de 2007

40. Leucemias,
http://www.drscope.com/pac/mg/a5/mga5_p23.htm
Fecha de Consulta: 30 de Marzo de 2007

41. Markiewicz T, Moszczyñski L. (2004) “Analysis of Features for Blood Cell Recognition.” VI International Workshop, Computational Problems of Electrical Engineering, Zakopane 2004.

42. Matlab user manual – Image processing toolbox, MathWorks, Natick, 1999

43. Medline Plus,
http://www.nlm.nih.gov/medlineplus/spanish/ency/article/003330.htm
Fecha de Consulta: 2 de Febrero de 2007

44. Modelo de Color CIE Lab, Marzo 23,2007
http://www.fotonostra.com/grafico/modelncs.htm
Fecha de Consulta: 23 de Marzo de 2007

45. Monteiro María-do-Céu, Martínez Marcial y O'connor José Enrique. “La citometría de flujo en el análisis funcional de las plaquetas: II. Aplicaciones clínicas”. Rev Diagn Biol. [online]. 2002, vol. 51, no. 3 [citado 2007-09-12], pp. 87-99

46. MSD,
http://www.msd.es/publicaciones/mmerck_hogar/seccion_14/seccion_14_152.html
Fecha de Consulta: 14 de Febrero de 2007

47. National Center for Health Statistics,
http://www.cdc.gov/nchs/howto/w2w/w2welcom.htm
Fecha de Consulta: 6 de Febrero de 2007

48. NewYork-Presbyterian Hospital,
http://wo-pub2.med.cornell.edu/cgi-bin/WebObjects/PublicA.woa/4/wa/viewHContent?website=nyp+spanish&contentID=4298&wosid=kGTI1YYSUp5VoLhJ3XnvDw
Fecha de Consulta: 22 de Febrero de 2007

49. Nipon Theera-Umpon “White Blood Cell Segmentation and Classification in Microscopic Bone Marrow Images”, L. Wang and Y. Jin (Eds.): FSKD 2005, LNAI 3614, pp. 787 – 796, 2005. Springer-Verlag Berlin Heidelberg 2005

50. Pajares G., M. de la Cruz J., *Vision by Computer,* AlfaOmega 1st ed, 123

51. Sánchez Segura Miriam, Rivero Jiménez René, Marsan Suárez Vianed et al. “Inmunofenotipaje en el diagnóstico de síndromes linfo y mieloproliferativos.” Rev Cubana Hematol Inmunol Hemoter, sep.-dic. 2000, vol.16, no.3, p.198-205.

52. Sanei Saeid, Lee Tracey, “Cell Recognition Based on PCA and Bayesian Classification”, 4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003), April 2003, Nara, Japan

53. SMOTE: Synthetic Minority Over-sampling Technique,
http://www.cs.cmu.edu/afs/cs/project/jair/pub/volume16/chawla02a-html/chawla2002.html
Fecha de Consulta: 23 de Marzo de 2007

54. Sucar L. Enrique, Gómez Gionavi, *Procesamiento de Imágenes y Visión Computacional*

55. Técnica de Análisis de Imagen,
http://books.google.com/books?id=Mi9UrJQCKoEC&pg=PA116&lpg=PA116&dq=imagen+numero+de+euler&source=web&ots=sPq-l3Knwm&sig=_GcYdp8Lw-XWDHcp5GQZbzbtoS8#PPA20-IA3,M1
Fecha de Consulta: 27 de Marzo de 2007

56. Técnicas Clásicas de Segmentación de Imagen,
http://poseidon.tel.uva.es/~carlos/ltif10001/segmenclasica.pdf
Fecha de Consulta: 2 de Febrero de 2007

57. The University of Chicago Medical Center,
http://www.uchospitals.edu/online-library/content=S03206

Fecha de Consulta: 12 de Febrero de 2007

58. Tomasz Markiewicz, Leszek Moszczyński, “Analysis of Features for Blood Cell Recognition”, VI International Workshop “Computational Problems of Electrical Engineering” Zakopane 2004

59. United States Cancer Statistics,
http://www.cdc.gov/cancer/npcr/npcrpdfs/US_Cancer_Statistics_2004_Incidence_and_Mortality.pdf
Fecha de Consulta: 19 de Febrero de 2007

60. Ventana al Universo,
http://www.windows.ucar.edu/earth/climate/images/spectrum_sm.sp.gif
Fecha de Consulta: 28 de Febrero de 2007

61. Visión Computacional,
http://turing.iimas.unam.mx/~elena/CompVis/Chap1-Intro.pdf.gz
Fecha de Consulta: 21 de Febrero de 2007

62. Visión por Computador,
http://www.depeca.uah.es/docencia/doctorado/cursos04_05/82854/docus/tr_vision(V).pdf
Fecha de Consulta: 16 de Marzo de 2007

63. WEKA,
http://www.cs.waikato.ac.nz/ml/weka/
Fecha de Consulta: 10 de Abril de 2007

64. WikiMEdia Commons,
http://commons.wikimedia.org/wiki/Image:Illu_blood_cell_lineage.jpg#file
Fecha de Consulta: 24 de Febrero de 2007

65. Wikipedia,
http://es.wikipedia.org/wiki/Sangre
Fecha de Consulta: 13 de Febrero de 2007

66. Wikipedia,
http://es.wikipedia.org/wiki/Imagen:AdditiveColorMixing.png
Fecha de Consulta: 13 de Febrero de 2007

67 La visión: Una ventana al mundo
http://www.icarito.cl/medio/articulo/0,0,38035857__264082087__1,00.html
Fecha de Consulta: 6 de Marzo de 2007

68. Segmentación: Umbralización, Regiones y Clústering
http://varpa.lfcia.org/Docencia/VAFiles/Curso0607/tema2.pdf
Fecha de Consulta: 6 de Marzo de 2007

69. Ensambles: Bagging, Boosting y variantes
http://ccc.inaoep.mx/~emorales/Cursos/KDD03/node53.html
Fecha de Consulta: 27 de Marzo de 2007

70 Aprendizaje Automático
http://scalab.uc3m.es/~docweb/aa/transpas/otros.pdf
Fecha de Consulta: 7 de Marzo de 2007