



Clasificación automática de células sanguíneas a partir de imágenes: comparación entre atributos GLCM y red neuronal convolucional



Introducción a la Ciencia de Datos Aplicada a Escenarios Médico-Biológicos
Semestre 2026-1

Frida Michelle Vargas Bautista

1. Introducción y Objetivo del Proyecto

1.1. Contexto general

En el ámbito de la salud, el análisis de muestras sanguíneas constituye una herramienta fundamental para la evaluación del estado fisiológico e inmunológico de un paciente. El reconocimiento y conteo de los distintos tipos de células presentes en la sangre periférica permite apoyar el diagnóstico, seguimiento y tratamiento de diversas enfermedades. Tradicionalmente, este análisis se realiza mediante la observación microscópica por parte de personal especializado, lo cual requiere tiempo, experiencia y recursos materiales.

En los últimos años, el desarrollo de técnicas de ciencia de datos, aprendizaje automático y visión de computadora ha abierto la posibilidad de automatizar parcialmente estos procesos, utilizando imágenes digitales de muestras sanguíneas. Estos enfoques buscan agilizar el análisis, reducir la carga de trabajo del personal clínico y ofrecer herramientas de apoyo para la toma de decisiones médicas, manteniendo un alto nivel de precisión.

1.2. Planteamiento del problema

A pesar de los avances en técnicas de aprendizaje automático, el reconocimiento automático de tipos celulares a partir de imágenes sanguíneas sigue siendo un reto, debido a la variabilidad morfológica entre células y a la similitud visual que puede existir entre distintos tipos celulares. Además, la forma en que se representan las imágenes para su análisis influye de manera directa en el desempeño de los modelos de clasificación.

En este contexto, surge la pregunta de si es posible identificar de manera confiable el tipo de célula sanguínea a partir de imágenes utilizando distintos enfoques de representación de los datos, como atributos texturales derivados de la imagen o el uso directo de las imágenes mediante modelos de aprendizaje profundo. Asimismo, resulta relevante evaluar qué tan informativa es cada representación y en qué medida influye en el desempeño de los modelos de clasificación.

1.3. Objetivo general

Desarrollar y evaluar modelos de clasificación supervisada para identificar el tipo de célula sanguínea a partir de imágenes, comparando el desempeño de enfoques basados en atributos texturales derivados de la matriz GLCM con un enfoque de aprendizaje profundo mediante una red neuronal convolucional.

1.4. Objetivos específicos

- Transformar las imágenes en datos estructurados mediante la extracción de atributos texturales utilizando la matriz de co-ocurrencia de niveles de gris (GLCM).
- Entrenar y evaluar distintos modelos de clasificación supervisada sobre los datos estructurados.
- Diseñar, entrenar y evaluar una red neuronal convolucional que utilice directamente las imágenes como entrada.
- Comparar el desempeño de los modelos basados en atributos tabulares y del enfoque convolucional utilizando métricas de evaluación apropiadas.

- Analizar las ventajas y limitaciones de cada enfoque en el contexto del reconocimiento automático de células sanguíneas.

2. Descripción del Dataset y Preparación de Datos

2.1. Origen y descripción del dataset

El conjunto de datos utilizado en este trabajo se obtuvo a partir de la plataforma *Kaggle*. Dicho *dataset* consta de un total de 17 092 imágenes de células sanguíneas individuales. Las imágenes fueron adquiridas mediante el analizador automatizado CellaVision DM96 en el Laboratorio Central del Hospital Clínic de Barcelona y se encuentran clasificadas por patólogos clínicos expertos. Las clases disponibles son: neutrófilos, eosinófilos, basófilos, linfocitos, monocitos, granulocitos inmaduros (ig), eritroblastos y plaquetas (o trombocitos).

De acuerdo con la descripción del *dataset*, las imágenes tienen un tamaño original de 360×363 píxeles, se encuentran almacenadas en formato JPG y corresponden a extendidos de sangre periférica. Además, se declara que las muestras provienen de individuos sanos, sin infección ni enfermedad hematológica u oncológica, y que no se encontraban bajo tratamiento farmacológico al momento de la recolección de la sangre.

Este *dataset* fue elegido por el hecho de que la clasificación celular haya sido realizada por expertos en la materia, lo cual nos da confianza de que se parte de bases sólidas. Adicionalmente, el hecho de que las imágenes provengan de pacientes sanos, sin condiciones hematológicas particulares, permite que los modelos extraigan las características más generales de cada tipo celular.

2.2. Calidad de datos y limpieza

Dado que se trata de un *dataset* público, se consideró relevante revisar tanto la documentación disponible como los comentarios reportados por la comunidad. En particular, en el repositorio de *Kaggle* algunos usuarios señalaron que las imágenes presentan tamaños heterogéneos, observación que fue corroborada mediante una inspección directa de los datos. A partir de ello, se desarrollaron funciones para estandarizar el tamaño de las imágenes a 128×128 píxeles, procurando no deformar su contenido.

Tipos de células sanguíneas en el DataSet

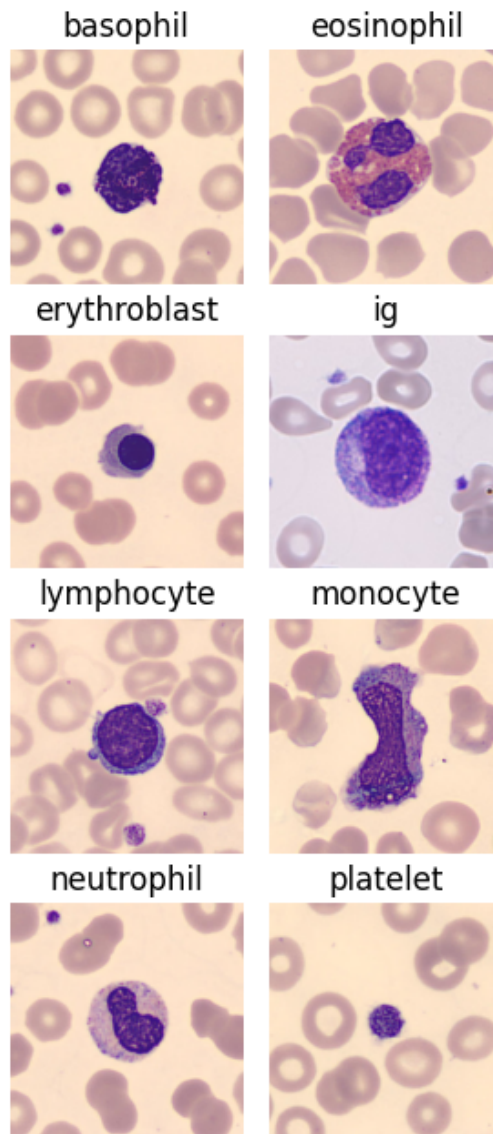


Figura 1: Ejemplos de diferentes células presentes en el *dataset*

Este reescalamiento se empleó únicamente en el preprocesamiento de las imágenes introducidas a la Red Neuronal Convolutiva (CNN). En contraste, para las imágenes que se transformaron a datos tabulares estructurados no fue necesario aplicar dicho reescalamiento, ya que al calcular la matriz de co-ocurrencia de niveles de gris (GLCM) se utilizó una versión normalizada de la matriz. En este caso, las entradas corresponden a probabilidades y no a frecuencias absolutas,

lo que permite comparar atributos entre imágenes de distinto tamaño. Más adelante se abordará con mayor detalle la interpretación de esta matriz y de la transformación asociada.

Respecto a posibles fuentes de ruido, como variaciones en la iluminación o en el enfoque de las imágenes, no se realizó una búsqueda exhaustiva de observaciones individuales anómalas. No obstante, en la construcción de la matriz GLCM se incorporó el parámetro de simetría, de modo que la matriz resultante es simétrica. Esta característica contribuye a que los atributos derivados sean más robustos frente a valores atípicos de intensidad. Además, dado que esta transformación captura variaciones locales entre píxeles vecinos, efectos como destellos de brillo tienden a producir cambios suaves en las intensidades dentro de un clúster, mientras que los saltos más pronunciados se concentran en las fronteras. Dichas variaciones se atenúan al promediarse sobre todas las vecindades posibles de la imagen, por lo que en un primer acercamiento se consideró que esta representación era suficiente para mitigar el impacto del ruido.

Cabe destacar que, para deducir la matriz GLCM asociada a cada imagen, fue necesario convertir previamente las imágenes a escala de grises, con el objetivo de preservar únicamente la intensidad de los píxeles y no la información específica de los canales RGB.

En el caso de las imágenes utilizadas como entrada directa a la red neuronal convolucional, las intensidades de los píxeles se dividieron entre 255 en cada uno de los tres canales (RGB), con el fin de escalar los valores a un rango común. Esta normalización evita que diferencias grandes en magnitud sesguen el proceso de entrenamiento y contribuye a un mejor desempeño de la CNN.

2.3. Transformaciones, selección de variables y dimensionalidad

Para convertir los datos no estructurados (imágenes) en datos estructurados en formato tabular se empleó la Matriz de Co-ocurrencia de Niveles de Gris (GLCM). Esta transformación parte de convertir la imagen a escala de grises, preservando únicamente la intensidad de cada píxel. Cada nivel de intensidad se representa mediante un número natural, cuyo rango depende de la escala empleada; en este trabajo se utilizó la escala de 0 a 255, donde el valor cero corresponde al color negro.

Dado que cada píxel tiene asignado un valor de intensidad, se compara con la intensidad de uno de sus píxeles vecinos —usualmente dentro de sus ocho veci-

nos inmediatos— y la transición entre ambos valores se registra en una matriz de frecuencias. Por ejemplo, si en la matriz la entrada correspondiente a la coordenada (234, 200) tiene valor 5, esto indica que existen cinco parejas de píxeles en la imagen donde el píxel de referencia tiene intensidad 234 y su vecino tiene intensidad 200.

Esta matriz permite analizar con qué frecuencia un píxel de cierta intensidad presenta vecinos con intensidades similares o significativamente distintas, lo que da información sobre la homogeneidad o heterogeneidad de la textura de la imagen. En particular, a partir de la matriz GLCM se pueden calcular diversos atributos, entre los cuales se emplearon los siguientes:

- **Contraste:** mide la variación local en la imagen. Valores altos indican mayores diferencias de intensidad entre píxeles vecinos.
- **Disimilaridad:** promedia la diferencia absoluta de intensidades entre píxeles vecinos; valores grandes sugieren una textura más heterogénea.
- **Homogeneidad:** relaciona la distribución de los elementos de la matriz con su diagonal. Valores altos indican que las transiciones de intensidad son pequeñas, lo que sugiere una textura más uniforme.
- **Energía:** valores altos están asociados con texturas más uniformes.
- **Correlación:** mide la dependencia lineal entre las intensidades de los píxeles vecinos.

Para el cálculo de la matriz y de estos atributos se empleó la librería *sklearn*, utilizando las funciones *graycomatrix* y *graycoprops*. En todos los casos, el píxel vecino considerado fue el píxel adyacente hacia la derecha.

Previo al cálculo de los atributos, se incorporaron dos especificaciones adicionales en la construcción de la matriz GLCM. La primera fue la normalización de la matriz, de modo que las entradas representan probabilidades en lugar de frecuencias absolutas. Por ejemplo, la entrada (234, 200) indica ahora la probabilidad de que un píxel con intensidad 234 tenga como vecino derecho un píxel con intensidad 200. Esta normalización permite comparar los atributos entre imágenes con diferente número de píxeles, ya que las relaciones se expresan de forma relativa al tamaño de la imagen.

La segunda especificación fue imponer simetría a la matriz. En general, no es esperable que las entradas (234, 200) y (200, 234) sean iguales, ya que la relación

depende de la dirección considerada. Sin embargo, al hacer la matriz simétrica se elimina esta dependencia direccional y se captura únicamente la relación de vecindad entre intensidades, lo cual aporta mayor robustez a la interpretación de los atributos calculados.

Otras transformaciones adicionales se abordarán en la siguiente sección, ya que su elección está fuertemente motivada por los resultados obtenidos durante el análisis exploratorio de datos.

Distribución de Atributos por Tipo de Célula, 1000 Muestras

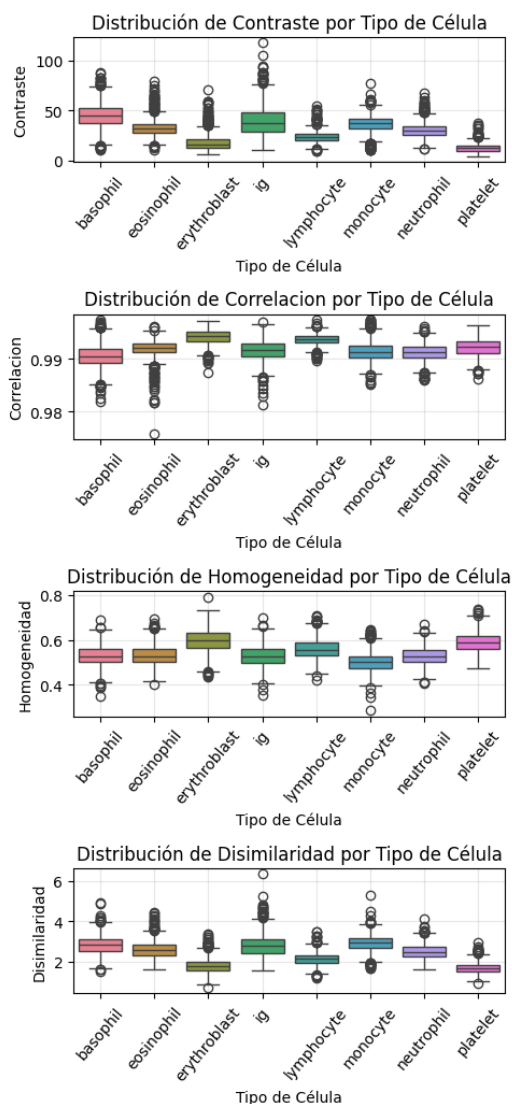


Figura 2: Diagramas de cajas y bigotes para atributos de la matriz GLCM, por clase.

3. Análisis Exploratorio de Datos (EDA)

3.1. Estadísticas descriptivas

El planteamiento inicial consistió en tomar una muestra aleatoria de 200 imágenes por clase del dataset y calcular, para cada una, los atributos derivados de su matriz GLCM. Con estos valores se construyeron diagramas de caja y bigotes con el objetivo de identificar los rangos, medianas y dispersión de las distribuciones por clase, así como explorar posibles diferencias significativas que permitieran discriminar entre tipos celulares.

Al graficar los diagramas correspondientes a las 1 600 imágenes iniciales, se observó una gran cantidad de valores atípicos. Con el propósito de evaluar si estos valores se debían a un tamaño de muestra reducido, se incrementó progresivamente el número de imágenes por clase, primero a 800 y posteriormente a 1 000. De manera interesante, el número de valores atípicos aumentó conforme creció la muestra, lo que sugiere que esta variabilidad extrema es intrínseca al problema y no producto de ruido aislado. En consecuencia, se consideró que imputar o eliminar estos valores podría resultar perjudicial, ya que, aunque atípicos, son influyentes.

Esta decisión motivó el uso de métricas y transformaciones menos sensibles a *outliers*. En particular, se empleó *RobustScaler()* de la librería *sklearn*, el cual, para cada variable, resta la mediana y divide entre el rango intercuartil. Este procedimiento permite colocar todas las variables en una escala comparable y reduce el riesgo de que clasificadores posteriores se vean sesgados por diferencias de magnitud.

Una vez escalados los datos, se exploró la posibilidad de identificar agrupamientos visuales. Dado que cada observación se encuentra en un espacio de cinco dimensiones —una por cada atributo de la matriz GLCM—, se evaluó la factibilidad de reducir la dimensionalidad a dos o tres dimensiones sin perder información relevante. Para ello, se aplicó un Análisis de Componentes Principales (PCA). El análisis de la varianza acumulada mostró que los dos primeros componentes explican cerca del 90 % de la variabilidad total, mientras que los tres primeros alcanzan aproximadamente el 97 %, valores considerados adecuados en la práctica.

Inicialmente se proyectaron los datos en el espacio bidimensional generado por los dos primeros componentes principales y se separaron las clases por color. No se observó una separación clara entre los distintos tipos celulares.

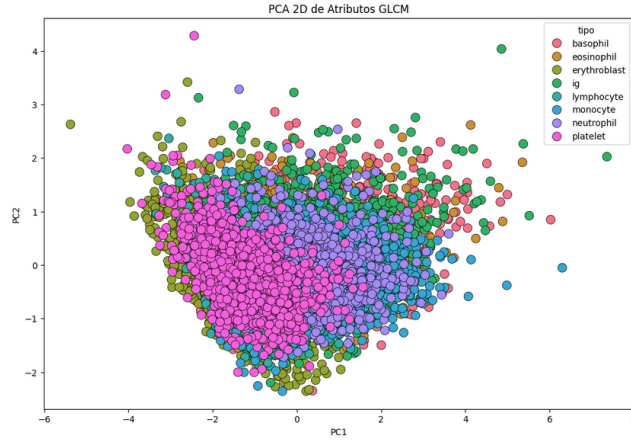


Figura 3: Proyección de los atributos de la matriz GLCM en el espacio de dos componentes principales

Posteriormente, se exploró la proyección en el espacio tridimensional definido por los tres primeros componentes principales, con el fin de identificar posibles diferencias en la distribución espacial. Aunque a primera vista no se apreciaron clústeres bien delimitados, un análisis más detallado reveló que la clase correspondiente a plaquetas presenta una mayor cohesión interna y una ubicación espacial más definida respecto a los demás tipos celulares.

Esta observación se profundizó al comparar visualmente la distribución de las plaquetas con la de un tipo celular adicional y, posteriormente, al agrupar todas las plaquetas en una clase y el resto de los tipos celulares en otra. En este escenario se observaron dos grupos diferenciados, aunque no perfectamente separados, lo que motivó la implementación de una Máquina de Soporte Vectorial (SVM) con kernel lineal.

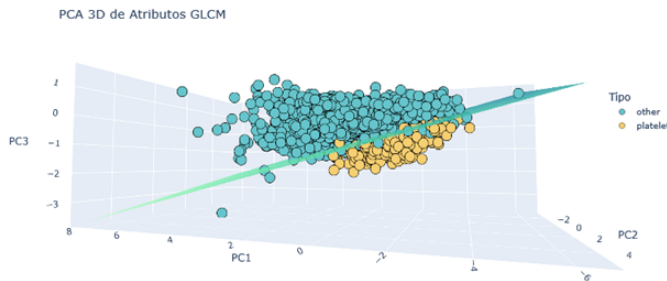


Figura 4: Datos etiquetados como "plaquetas" y "otro" tipo celular sanguíneo. También se grafica la SVM ajustada para separar estas dos clases.

La elección de una SVM, en lugar de un algoritmo como *k-means*, se debió a la forma de la nube de pun-

tos: mientras que *k-means* busca minimizar distancias a centroides, la SVM se enfoca en maximizar el margen entre las fronteras de ambas clases, lo cual resulta más adecuado para distribuciones con geometría elipsoidal.

El plano ajustado mostró un desempeño visualmente razonable, lo que sugiere que vale la pena explorar, en trabajos futuros, el uso de SVM con kernels más complejos, como polinomial o RBF. No obstante, dado que el objetivo principal de este trabajo es la clasificación multiclase, este análisis se reporta únicamente como un hallazgo exploratorio. Resulta especialmente interesante desde el punto de vista biológico indagar por qué las imágenes de plaquetas presentan atributos texturales más diferenciados.

Finalmente, dado que no se identificaron clústeres bien definidos entre los ocho tipos celulares al reducir la dimensionalidad, se procedió a un enfoque supervisado. Se retomó la tabla de datos escalada, conservando los cinco atributos originales, y se entrenaron distintos algoritmos de clasificación supervisada, evaluando su desempeño sobre un subconjunto de prueba. De manera adicional, se implementó una red neuronal convolucional para realizar la misma tarea de clasificación directamente sobre las imágenes y comparar su desempeño con los modelos entrenados sobre datos tabulares.

3.2. Insights principales del EDA

Los principales *insights* obtenidos durante el análisis exploratorio de datos se resumen a continuación:

- Las imágenes constituyen un tipo de dato no estructurado, mientras que muchas técnicas clásicas de clasificación supervisada están diseñadas para datos estructurados. Esto motiva el uso de transformaciones, como la matriz GLCM, que permiten extraer atributos y analizarlos en formato tabular.
- A partir de los atributos derivados de la matriz GLCM, los diagramas de caja y bigotes muestran que las medianas entre clases se concentran en valores similares y que los rangos intercuartiles son amplios, lo que dificulta identificar atributos exclusivos o altamente representativos de cada tipo celular. Esto sugiere la necesidad de emplear algoritmos de clasificación que puedan manejar solapamientos entre clases.
- Al reducir el espacio de atributos a tres dimensiones mediante PCA, se observa que la clase correspondiente a plaquetas presenta una mayor cohesión interna y una distribución espacial diferenciada respecto a los demás tipos celulares, lo cual

representa un hallazgo interesante tanto desde el punto de vista de clasificación como desde una posible interpretación biológica.

4. Modelado de Datos

4.1. Formulación del problema de modelado

El problema abordado en este trabajo se formula como un problema de clasificación supervisada multiclase, cuyo objetivo es predecir el tipo de célula sanguínea presente en una imagen a partir de información visual.

Cada observación corresponde a una imagen de una célula sanguínea etiquetada en una de las ocho clases disponibles en el dataset. Con el fin de evaluar distintas estrategias de representación, se consideran dos enfoques para abordar el mismo problema de clasificación. El primero utiliza una representación estructurada basada en atributos texturales derivados de la matriz de co-ocurrencia de niveles de gris (GLCM). El segundo emplea directamente las imágenes originales como entrada a una red neuronal convolucional, permitiendo que el modelo aprenda representaciones jerárquicas a partir de los datos no estructurados.

La variable objetivo es una variable categórica multiclase que indica el tipo celular. Dado que se trata de un problema supervisado, el conjunto de datos se divide en subconjuntos de entrenamiento y prueba, cuidando que la información del conjunto de prueba no influya en el ajuste de los modelos.

Durante el análisis exploratorio de datos se observó una alta variabilidad intrínseca dentro de cada clase y una superposición considerable entre distintos tipos celulares en el espacio de atributos. Estas características justifican la comparación entre modelos clásicos entrenados sobre atributos derivados y un enfoque de aprendizaje profundo, con el objetivo de evaluar si el uso directo de las imágenes permite capturar patrones más complejos relevantes para la clasificación.

El desempeño de ambos enfoques se evalúa en función de su capacidad para discriminar correctamente entre los distintos tipos celulares, utilizando métricas apropiadas para problemas de clasificación multiclase.

4.2. Modelos utilizados y justificación

El enfoque metodológico seguido en este trabajo se basa en explicaciones y recomendaciones proporcionadas por el profesor durante el curso, así como en material educativo complementario.

Se emplearon diversos algoritmos de clasificación supervisada multiclase, con el objetivo de explorar distintas estrategias de modelado y evaluar cuáles presentan un mejor desempeño en la identificación de los tipos celulares. En particular, se consideraron ensambles de árboles de decisión, Máquinas de Soporte Vectorial (SVM) con diferentes kernels y un modelo de regresión logística.

Un *ensamble* es un enfoque de *Machine Learning* en el que se combinan los resultados de varios modelos sencillos con el fin de construir un modelo más robusto y mejorar su desempeño predictivo. Dos de las estrategias más comunes para construir ensambles son el *bagging* y el *boosting*. En el caso del *bagging*, cada modelo base se entrena de manera independiente sobre distintos subconjuntos del conjunto de entrenamiento. Este es el principio detrás de los **Random Forest**, los cuales se construyen a partir de múltiples árboles de decisión. Cada árbol se entrena con una muestra diferente del dataset y genera su propia predicción; al momento de clasificar una observación, cada árbol emite un veredicto y la clase mayoritaria es devuelta como la predicción final.

Por otro lado, la técnica de *boosting* entrena los modelos de forma iterativa, de tal manera que cada nuevo modelo se enfoca en corregir los errores cometidos por los anteriores. En este proceso, se asigna mayor peso a las observaciones mal clasificadas en etapas previas, permitiendo que los modelos subsecuentes se concentren en ellas y mejoren progresivamente el desempeño global. Dos algoritmos basados en ensambles de árboles que utilizan este enfoque son **Gradient Boosting** y **XGBoost**.

Con el fin de hacer comparables estos tres métodos de ensamble, se fijó el parámetro *n_estimators* en 100 para todos ellos. Este parámetro indica el número de árboles que conforman el ensamble y se eligió como un valor estándar que ofrece un equilibrio razonable entre estabilidad del modelo y costo computacional.

Adicionalmente, se exploró el uso de Máquinas de Soporte Vectorial (SVM) con distintos kernels: lineal, polinomial y de base radial (RBF). Las SVM buscan encontrar una frontera de decisión que maximice la separación entre las distintas clases, y la elección del kernel determina la forma de dicha frontera. El kernel lineal ajusta un plano o hiperplano, dependiendo de la dimensionalidad del espacio de atributos; en secciones anteriores se mostró, por ejemplo, cómo este modelo ajusta un plano en un espacio tridimensional. El kernel polinomial utiliza funciones polinomiales como base, mientras que el kernel RBF se apoya en una base de dimensión infinita. Estos dos últimos resultan particu-

larmente relevantes para modelar relaciones no lineales entre las variables.

Por otra parte, se ajustó un modelo de regresión logística multiclase, el cual corresponde a un clasificador lineal que, en la etapa de predicción, devuelve un conjunto de probabilidades asociadas a la pertenencia de cada observación a las distintas clases. Este modelo se utilizó principalmente como un punto de referencia frente a modelos de mayor complejidad.

Finalmente, se entrenó una red neuronal convolucional (CNN) que recibe como entrada imágenes de tamaño 128×128 con intensidades normalizadas en el intervalo $[0, 1]$. La arquitectura empleada consta de tres capas convolucionales con 32, 64 y 128 filtros, respectivamente, todas ellas con función de activación *ReLU*. Se utilizó *padding* para preservar el tamaño espacial de las imágenes a lo largo de las capas convolucionales, evitando que se reduzca progresivamente y permitiendo que las capas densas posteriores dispongan de mayor información.

Posteriormente, se incorporaron dos capas densas con función de activación *ReLU*, donde en la primera se realiza el aplanamiento de la salida convolucional mediante la operación *flatten*. En la última capa se utilizó la función de activación *softmax*, ya que el objetivo es obtener probabilidades de pertenencia a cada una de las clases. Durante el entrenamiento se empleó un tamaño de lote (*batch size*) normalizado, lo cual, si bien introduce cierta discontinuidad en el gradiente utilizado para la optimización, ayuda a mantener una propagación más estable de la información entre capas y a reducir posibles sesgos durante el aprendizaje.

A este punto, vale la pena mencionar dos consideraciones importantes. La primera es que no se realizó una optimización exhaustiva de hiperparámetros. Esta decisión se tomó de manera deliberada, ya que el objetivo principal del estudio fue comparar el desempeño de distintos modelos de clasificación bajo condiciones equivalentes. De esta forma, además de facilitar una comparación justa entre algoritmos de naturaleza distinta, se obtiene información indirecta sobre qué tan representativos son los atributos derivados de la matriz GLCM para identificar el tipo celular.

En particular, si modelos con arquitecturas y supuestos diferentes presentan desempeños similares, esto sugiere que la información contenida en los atributos es consistente y que, independientemente del clasificador utilizado, estos describen de manera semejante a los datos. Por otro lado, si todos los modelos muestran un desempeño bajo, esto indicaría que los atributos derivados de la matriz GLCM no son suficientes para capturar las diferencias entre clases. Este escenario

abriría distintas líneas de acción, como la incorporación de un mayor número de atributos texturales, la exploración de otras técnicas para transformar datos no estructurados en representaciones tabulares, o bien el uso directo de modelos de aprendizaje profundo, como redes neuronales convolucionales, desde etapas tempranas del análisis.

Bajo esta perspectiva, el presente estudio evalúa de manera indirecta la calidad y capacidad descriptiva de los datos transformados mediante la matriz GLCM. Por esta razón, no se realizaron ajustes de hiperparámetros en esta etapa. No obstante, se reconoce que la implementación de búsquedas de hiperparámetros, ya sea mediante *Grid Search* o, de manera más eficiente, *Random Search*, constituye una línea natural de trabajo futuro para analizar si un ajuste fino de los modelos puede mejorar significativamente su desempeño.

4.3. Estrategia de entrenamiento y validación

Para el caso de los datos estructurados, se seleccionaron de manera aleatoria 1 000 imágenes por clase a partir del dataset original. De estas, 800 imágenes se destinaron al conjunto de entrenamiento y 200 al conjunto de evaluación. Dado que todas las clases están igualmente representadas, esta partición garantiza un balance adecuado entre clases y evita sesgos derivados de desbalance en la variable objetivo.

Con el fin de reducir el riesgo de sobreajuste y obtener una estimación más robusta del desempeño de los modelos, se empleó validación cruzada con 5 particiones (*5-fold cross validation*) sobre el conjunto de entrenamiento. En este esquema, los datos de entrenamiento se dividen en cinco subconjuntos, de modo que en cada iteración cuatro de ellos se utilizan para entrenar el modelo y el restante para validarlo, rotando esta asignación entre los folds. El desempeño reportado corresponde al promedio obtenido a lo largo de las distintas particiones.

El conjunto de evaluación se mantuvo completamente separado durante el proceso de entrenamiento y validación cruzada, y se utilizó únicamente para estimar el desempeño final de los modelos, evitando así cualquier forma de contaminación entre entrenamiento y prueba.

En el caso de los datos utilizados para entrenar la red neuronal convolucional, el conjunto total de imágenes se dividió en tres subconjuntos: entrenamiento, validación y prueba. La partición se realizó de manera aleatoria siguiendo una proporción aproximada de 70 %

para entrenamiento, 15 % para validación y 15 % para prueba.

Con el objetivo de mejorar la capacidad de generalización del modelo se emplearon técnicas de *data augmentation*. En particular, se aplicaron transformaciones geométricas simples sobre las imágenes del conjunto de entrenamiento, tales como volteos horizontales y verticales. Estas transformaciones permiten generar variaciones realistas de las imágenes sin alterar la identidad de la célula.

Durante el entrenamiento se utilizó un tamaño de lote (*batch size*) de 32, lo que indica el número de

imágenes procesadas de manera simultánea en cada iteración del algoritmo de optimización. Inicialmente se entrenó el modelo durante tres épocas (*epochs*), donde una época corresponde a un recorrido completo por el conjunto de entrenamiento. Sin embargo, se observó un sobreajuste marcado, evidenciado por una mejora en el desempeño sobre el conjunto de entrenamiento sin una mejora correspondiente en el conjunto de validación. En consecuencia, se redujo el número de épocas a dos, buscando limitar el sobreajuste y favorecer una mejor generalización del modelo.

5. Resultados y Evaluación del Modelo

5.1. Resultados cuantitativos

	modelo	accuracy	f1_micro	f1_macro	balanced_accuracy	precision_macro	recall_macro	jaccard_macro
5	SVM RBF	0.607344	0.607344	0.603009	0.607423	0.609222	0.607423	0.456331
4	SVM Linear	0.601250	0.601250	0.596393	0.601376	0.598547	0.601376	0.453416
0	Random Forest	0.591406	0.591406	0.587946	0.591363	0.588377	0.591363	0.442026
2	XGBoost	0.580937	0.580937	0.579434	0.580955	0.580018	0.580955	0.434656
1	Gradient Boosting	0.581562	0.581562	0.576546	0.581827	0.578538	0.581827	0.431177
6	Logistic Regression	0.581406	0.581406	0.572158	0.581417	0.571155	0.581417	0.429862
3	SVM Poly	0.525625	0.525625	0.536223	0.526903	0.599608	0.526903	0.393465

Figura 5: Tabla comparando el desempeño de diversos algoritmos de clasificación multiclase, empleando métricas variadas.

Se evaluó el desempeño de los distintos modelos de clasificación utilizando varias métricas, y para la comparación final se tomó como referencia principal el *F1-score*, ya que combina de manera equilibrada la *precision* y la sensibilidad (*recall*).

Entre los clasificadores entrenados sobre datos estructurados, el modelo que presentó el mejor desempeño fue la Máquina de Soporte Vectorial (SVM) con kernel RBF, alcanzando un *F1-score* aproximado del 60 %. Para este clasificador se calculó la matriz de confusión sobre el conjunto de prueba. A partir de ella se observó que las clases que el modelo identifica con mayor precisión son las plaquetas y los basófilos, mientras que las mayores confusiones se presentan en las clases de eosinófilos e Ig.

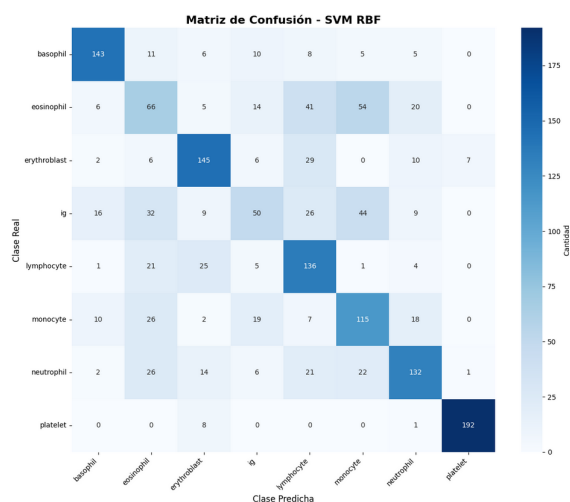


Figura 6: Matriz de confusión asociada a la predicción de la SVM con kernel RBF, sobre los datos de prueba.

En el caso de la red neuronal convolucional, el entrenamiento se realizó optimizando la *accuracy* sobre el conjunto de validación. Se empleó un criterio de *early stopping*, mediante el cual el modelo conserva los pesos correspondientes al mejor desempeño observado en el conjunto de validación. En particular, si la *accuracy* de validación no mejora durante dos épocas consecutivas, el entrenamiento se detiene y se restauran los últimos pesos correspondientes al mejor modelo.

Bajo este esquema, el entrenamiento se detuvo en la segunda época, la cual presentó el mejor desempeño en términos de *accuracy* sobre el conjunto de validación.

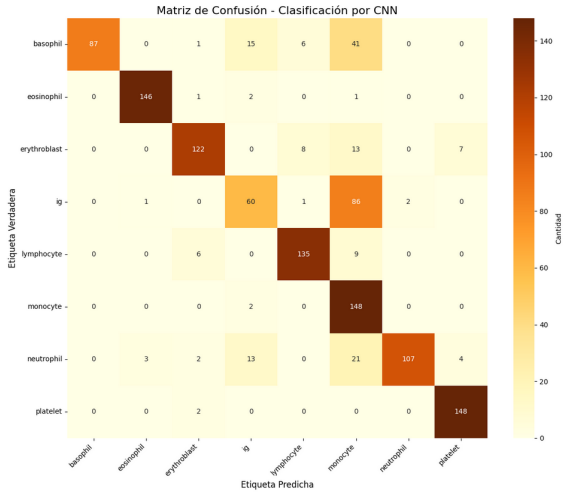


Figura 7: Matriz de confusión asociada a la predicción de la CNN con kernel RBF, sobre los datos de prueba.

Además su *f1-score macro* es del 79%.

5.2. Métricas de clasificación o regresión

Para evaluar el desempeño de los distintos modelos de clasificación multiclase se reportaron diversas métricas complementarias, entre ellas *accuracy*, *balanced accuracy*, *precision*, *recall*, *F1-score* y el índice de Jaccard, en sus versiones *micro* y *macro*. Cada una de estas métricas captura un aspecto distinto del comportamiento de los clasificadores, por lo que su análisis conjunto permite una evaluación más completa que el uso de una sola medida.

La *accuracy* mide la proporción total de predicciones correctas sobre el conjunto de evaluación. Si bien es una métrica intuitiva, puede resultar limitada en problemas de clasificación multiclase, ya que no distingue si los errores se concentran en clases específicas. Por

esta razón, también se reportó la *balanced accuracy*, la cual promedia la tasa de aciertos por clase y permite una evaluación más equitativa del desempeño global del modelo.

La *precision* cuantifica qué proporción de las observaciones que el modelo asigna a una clase son efectivamente correctas, mientras que el *recall* mide qué proporción de las observaciones reales de una clase fue correctamente identificada. Ambas métricas son relevantes cuando se desea analizar el tipo de errores cometidos por el clasificador. El *F1-score* combina *precision* y *recall*, penalizando escenarios en los que una de ellas es alta y la otra baja, por lo que proporciona una medida equilibrada del desempeño del modelo.

En este trabajo, el *F1-score* se utilizó como la métrica principal para la comparación entre modelos, ya que ofrece un balance adecuado entre *precision* y *recall* y resulta más informativo que la *accuracy* en un problema de clasificación multiclase donde varias clases presentan proporciones similares de aciertos y errores. Dado que el conjunto de datos se encuentra balanceado entre clases, las versiones *micro* y *macro* del *F1-score* presentan valores cercanos. No obstante, la versión *macro* evalúa el desempeño de manera independiente para cada clase y posteriormente promedia los resultados, ignorando el tamaño absoluto de cada una y enfocándose únicamente en las proporciones relativas. Por esta razón, se considera una medida más estricta y robusta del desempeño global del modelo.

Finalmente, se reportó también el índice de Jaccard, el cual mide la similitud entre el conjunto de etiquetas predichas y el conjunto de etiquetas reales.

5.3. Análisis crítico de resultados

En términos globales, la red neuronal convolucional presenta un mejor desempeño cuantitativo que los modelos entrenados sobre datos estructurados, con una diferencia aproximada de 20 puntos porcentuales en el *F1-score* respecto al mejor clasificador basado en atributos GLCM. No obstante, un análisis más detallado a través de las matrices de confusión revela comportamientos complementarios entre ambos enfoques que no se reflejan completamente en las métricas agregadas.

En particular, se observa que el clasificador basado en datos estructurados (SVM con kernel RBF) muestra un mejor desempeño en la identificación de ciertas clases específicas, como los basófilos, donde presenta menos confusiones en comparación con la red neuronal convolucional. En contraste, la CNN logra una clasificación considerablemente más precisa en clases como los eosinófilos, para las cuales los modelos basados en

atributos texturales presentan mayores dificultades.

Estas diferencias sugieren que cada enfoque captura distintos aspectos de la información contenida en las imágenes. Mientras que los modelos basados en la matriz GLCM parecen aprovechar mejor ciertos patrones texturales globales, la red neuronal convolucional es más efectiva para aprender representaciones visuales complejas directamente de los datos, lo que resulta especialmente ventajoso para algunas clases celulares.

A partir de esta observación, se abren líneas interesantes de trabajo futuro, como la exploración de esquemas de ensamble que combinen ambos enfoques, asignando pesos diferenciados por clase en función de su desempeño individual. Este tipo de estrategias podría permitir aprovechar las fortalezas de cada método, aunque sería necesario estudiar cuidadosamente su impacto sobre la generalización del modelo y el riesgo de sobreajuste.

6. Conclusiones y Trabajo a Futuro

6.1. Conclusiones principales

En este trabajo se abordó el problema de la clasificación automática de células sanguíneas normales a partir de imágenes, explorando dos enfoques distintos de representación de los datos: modelos de clasificación supervisada entrenados sobre atributos texturales derivados de la matriz GLCM y un enfoque de aprendizaje profundo mediante una red neuronal convolucional.

Los resultados muestran que los modelos basados en atributos estructurados presentan un desempeño limitado, con valores moderados de *F1-score*, lo cual sugiere que los atributos extraídos mediante la matriz GLCM no son suficientes para separar de manera clara a todas las clases celulares. Este comportamiento es consistente con el análisis exploratorio de datos, donde se observó que varias clases comparten valores similares en dichos atributos.

Por otro lado, la red neuronal convolucional alcanzó un mejor desempeño global, evidenciando su capacidad para aprender representaciones más complejas directamente a partir de las imágenes. No obstante, el análisis de las matrices de confusión reveló que ambos enfoques presentan fortalezas distintas dependiendo de la clase celular, lo que indica que ninguno de los métodos captura por completo toda la variabilidad presente en los datos.

En conjunto, estos resultados destacan la importancia de la representación de los datos en problemas de

clasificación de imágenes biomédicas y muestran que el uso de modelos más complejos no siempre garantiza un desempeño uniforme en todas las clases.

6.2. Limitaciones del estudio

Una de las principales limitaciones de este estudio es que los modelos entrenados sobre datos estructurados dependen de un conjunto reducido de atributos texturales derivados de la matriz GLCM. Si bien estos atributos capturan información relevante de la textura, no incorporan otros aspectos morfológicos importantes de las células, como la forma, el tamaño o la relación núcleo-citoplasma, lo cual puede limitar su capacidad discriminativa. Otra limitación inherente de la matriz GLCM es que el vecino con el que se compara cada píxel es sólo uno, lo que puede reducir la captura de información más global de la imagen.

Asimismo, aunque la red neuronal convolucional mostró un mejor desempeño global, su entrenamiento se realizó sobre un subconjunto balanceado del dataset original, lo que pudo afectar su capacidad de generalización. Además, no se llevó a cabo una optimización exhaustiva de hiperparámetros, ya que el enfoque del trabajo se centró en la comparación metodológica entre representaciones más que en maximizar el desempeño absoluto de los modelos.

Finalmente, este estudio se limita a la clasificación de células sanguíneas normales, por lo que los resultados no pueden extrapolarse directamente a escenarios clínicos que involucren patologías hematológicas.

6.3. Líneas del trabajo

Este trabajo se enmarca en las líneas de clasificación de imágenes biomédicas, transformación de datos no estructurados, comparación de modelos de clasificación supervisados y visión por computadora.

6.4. Líneas de trabajo futuro

A partir de los resultados obtenidos, se identifican diversas líneas de trabajo futuro, ya que por cada decisión metodológica fue posible considerar alternativas que podrían producir resultados igualmente interesantes. Por ejemplo, se podrían explorar otras transformaciones distintas a la matriz GLCM para convertir las imágenes en datos tabulares. En el caso de continuar con GLCM, sería posible experimentar con el número y la ubicación de los vecinos utilizados para comparar cada píxel, así como evaluar si la inclusión de un mayor

número de atributos derivados de la matriz mejora de manera significativa el desempeño de los clasificadores.

Otra línea relevante consiste en realizar una búsqueda sistemática de hiperparámetros tanto para el mejor modelo de clasificación de datos estructurados como para la red neuronal convolucional. Asimismo, dado que las plaquetas resultaron ser el grupo mejor diferenciado por ambos enfoques, podría explorarse una estrategia jerárquica en la que primero se identifiquen las plaquetas y posteriormente se clasifiquen los demás tipos celulares, con el objetivo de reducir costos computacionales o mejorar la atención del modelo sobre las clases restantes.

Adicionalmente, dado que para la construcción de la matriz GLCM las imágenes se transformaron a escala de grises, se pierde información relacionada con el color. Por ello, podría explorarse el uso de transformaciones que incorporen información cromática. Otra posible línea de investigación es aplicar el uso de la matriz GLCM sobre imágenes de tejidos con cáncer, donde la textura del tejido sí presenta cambios más marcados en presencia de agrupaciones celulares malignas.

Finalmente, los resultados sugieren la posibilidad de desarrollar esquemas de ensamble que combinen modelos basados en atributos estructurados y modelos convolucionales, asignando pesos diferenciados por clase con el objetivo de aprovechar las fortalezas de cada enfoque. Este tipo de aproximaciones podría contribuir a mejorar el desempeño global del sistema, aunque requeriría un análisis cuidadoso para evitar problemas de sobreajuste.

Referencias

- UncleSamulus. (2018). *Blood cells image dataset*. Kaggle. <https://www.kaggle.com/datasets/unclesamulus/blood-cells-image-dataset>
- Chollet, F., et al. (2015). *Keras*. <https://keras.io>
- DataCamp. (s.f.). *Introduction to Deep Learning with Keras*. <https://www.datacamp.com>
- DataCamp. (s.f.). *Biomedical Image Analysis in Python*. <https://www.datacamp.com>
- Hall-Beyer, M. (2017). *GLCM texture: A tutorial* (Version 3.0). University of Calgary. <https://lapi.fi-p.unam.mx/assets/img/materias/RP/materialesPracticas/Tutorial-GLCM.pdf>

7. Reproducibilidad y Repositorio en GitHub

7.1. Detalles del repositorio

En el repositorio se encuentra, además del PDF del presente reporte, una copia del notebook de Google Colab donde se desarrollaron todas las implementaciones y pruebas realizadas, así como un archivo *README* con una descripción general del trabajo.

Se incluye la url del repositorio: <https://shorturl.at/7F6yb>. La limitación es que es una liga abreviada y su acceso al repositorio se rompe si no es usada al menos una vez al mes por lo que si no funciona se sugiere escanear el QR.

7.2. Código QR del repositorio de GitHub

Empleando la página web de *The QR generator* se realizó el siguiente código QR donde al escanearlo dirigirá a la página del repositorio.



Figura 9: Código QR para el repositorio