

Universidad Nacional de San Agustín
Facultad de Ingeniería de Producción y Servicios
Escuela Profesional de Ciencia de la Computación



Exploración de *Vision Transformer* para la clasificación de células normales de sangre periférica

Docente:
Dr. Juan Carlos Gutierrez Caceres

Presentado por:
Abel Edmundo Borit Guitton
Luis Alberto Borit Guitton
Betzy Jacqueline Yarín Ramírez

**Arequipa - Perú
2024**

ÍNDICE

I.	Introducción	3
I-A.	Descripción General	3
I-B.	Justificación	3
I-C.	Objetivos	4
II.	Marco Teórico	4
II-A.	Fundamentos de la Hematología y la importancia del análisis de la sangre periférica	4
II-A1.	La sangre	4
II-A2.	Componentes de la sangre	4
II-A3.	Importancia del análisis sanguíneo	5
II-B.	Modelos propuestos para la clasificación de células sanguíneas	6
II-B1.	Redes Convolucionales (<i>CNNs</i>)	6
II-B2.	<i>Vision Transformer (ViT)</i>	6
III.	Diseño e Implementación	7
III-A.	Dataset	7
III-B.	Librerías y paquetes	8
III-C.	Diseño	8
III-C1.	Estrategia de Implementación	8
III-C2.	Preparación de datos	9
III-C3.	Definición del Modelo	9
III-C4.	Entrenamiento del Modelo	9
III-C5.	Evaluación del Modelo	9
III-C6.	Visualización de Resultados	9
III-D.	Implementación	9
IV.	Resultados y Discusión	10
V.	Conclusiones	13
	Referencias	13

ÍNDICE DE FIGURAS

1.	Imagen inmunoglobulina del <i>Dataset</i>	5
2.	Imagen eritroblasto del <i>Dataset</i>	5
3.	Eosinófilo, basófilo y neutrófilo del <i>Dataset</i>	5
4.	Monocito y linfocito del <i>Dataset</i>	5
5.	Imagen de plaquetas del <i>Dataset</i>	5
6.	Arquitectura de <i>CNN</i>	6
7.	Ilustración <i>Transformer Encoder</i> fue inspirada por Vaswani et al. (2017)	7
8.	Clasificación de células sanguíneas con predicciones de probabilidad	9
9.	Matriz de confusión del rendimiento	11
10.	Imagen clasificada como Basófilo	11
11.	Imagen clasificada como Eosinófilo	11
12.	Imagen clasificada como Eritroblasto	11
13.	Imagen clasificada como Linfocito	11
14.	Imagen clasificada como Monocito	12
15.	Imagen clasificada como Neutrófilo	12
16.	Imagen clasificada como Plaquetas	12
17.	Imagen clasificada como Ig	12

Exploración de visión transformer para la clasificación de células normales de sangre periférica

Maestría en Ciencias de la Computación, Universidad Nacional de San Agustín

1st Abel Edmundo Borit Guitton
Participante de Maestría
Arequipa, Perú
aborit@unsa.edu.pe

2nd Luis Alberto Borit Guitton
Participante de Maestría
Arequipa, Perú
lborit@unsa.edu.pe

3rd Betzy Jacqueline Yarín Ramírez
Participante de Maestría
Arequipa, Perú
byarin@unsa.edu.pe

I. INTRODUCCIÓN

I-A. Descripción General

El análisis de células sanguíneas es una parte importante de la evaluación de la salud y la inmunidad. El recuento y la densidad de estas células sanguíneas se utilizan para encontrar múltiples trastornos como infecciones de la sangre (anemia, leucemia, entre otras). Los métodos tradicionales sigue siendo una práctica común, a pesar de ser un proceso laborioso y que requiere mucho tiempo. [1] Sin embargo, esta tarea realizada por los profesionales de la salud está sujeta a errores y variaciones, lo que puede conducir a inconsistencias en los resultados. Además, el aumento en el volumen de muestras a procesar incrementa la carga de trabajo para los profesionales, lo que subraya la necesidad de soluciones precisas, eficientes y rápidas.

Para abordar este desafío, la comunidad científica está inmersa en un continuo desarrollo de nuevas tecnologías automatizadas, apoyadas por la inteligencia artificial. Especialmente, las redes neuronales convolucionales (*CNNs*) han demostrado mejoras significativas en la precisión y velocidad del diagnóstico. Actualmente, las arquitecturas de *Vision Transformer (ViT)* están captando atención debido a su capacidad para ofrecer resultados similares a las *CNNs*, pero con un tiempo de procesamiento menor.

Estos avances tienen como objetivo mejorar la eficiencia y rapidez del diagnóstico, reducir la variabilidad entre observadores y aumentar la capacidad de procesamiento de las muestras, lo que resultaría en una atención médica más efectiva y oportuna. Esto beneficia tanto a los pacientes como a los profesionales de la salud.

I-B. Justificación

El desarrollo de tecnologías de reconocimiento de imágenes ha experimentado avances significativos en estos últimos años, especialmente con el surgimiento de arquitecturas basadas en redes neuronales profundas (*Deep Learning*).

Previamente, las arquitecturas más comunes se basaban en *Convolutional Neural Networks (CNNs)*, las cuales han demostrado ser eficaces en una variedad de tareas de visión por computadora. Sin embargo, las *CNN* dependen en gran medida de campos receptivos locales y operaciones de agrupación, lo que impone limitaciones a su capacidad para capturar dependencias de largo alcance dentro de una imagen. Esta restricción impide su potencial para adquirir una comprensión integral de la entrada y comprender relaciones intrincadas entre varias regiones de la imagen. Además, las arquitecturas *CNN* con frecuencia requieren esfuerzos meticulosos de ingeniería y optimización para lograr un rendimiento óptimo en conjuntos de datos particulares, lo que reduce su flexibilidad y adaptabilidad. En cambio las arquitecturas basadas en *Visual Transformer (ViT)*, tienen varias ventajas que las hacen más atractiva, ofrecen una arquitectura más general y universal. [2]

Una de las principales ventajas de utilizar *ViT* en comparación con las arquitecturas anteriores, como *CNNs*, radica en su capacidad para capturar relaciones de largo alcance entre píxeles en una imagen. Mientras que las *CNNs* se basan en la convolución local para extraer características, *ViT* utiliza una atención global a través de transformadores, lo que permite capturar patrones más complejos en las imágenes. Esto es muy beneficioso para aplicaciones médicas donde la morfología celular puede variar significativamente y donde la captura de características a diferentes escalas y contextos es esencial.

Adicionalmente, *ViT* ofrece más flexibilidad y escalabilidad, permitiendo el procesamiento de imágenes con alta resolución. Esto es fundamental en el reconocimiento de células sanguíneas, donde la calidad y la resolución de las imágenes pueden variar según la técnica de adquisición y la calidad de la muestra.

En este trabajo, se propone estudiar la aplicación de *ViT* para el reconocimiento de células en sangre periférica con el objetivo de comprobar su capacidad para resolver esta problemática con alta precisión. Se espera que los resultados

de este estudio proporcionen información valiosa sobre el potencial de *ViT* en aplicaciones médicas específicas.

I-C. Objetivos

Objetivo General:

- El propósito de este proyecto es emplear un modelo basado en *Vision Transformer* (*ViT*) para identificar automáticamente las células normales de sangre periférica, a partir de un conjunto de imágenes.

Objetivos Específicos:

- Entender el proceso de reconocimiento de distintos tipos de células sanguíneas periféricas en imágenes digitales.
- Ampliar los conocimientos en Deep Learning, especialmente en relación con *ViT*.
- Desarrollar y validar modelos de detección automática basados en *CNNs* y *ViT* utilizando imágenes digitales de células sanguíneas periféricas.
- Comparar el rendimiento y la efectividad del modelo con los algoritmos convencionales basados en *CNNs*.

II. MARCO TEÓRICO

En las últimas décadas, el campo de la Visión Artificial ha atravesado una transformación radical en el sector sanitario, impulsada por el avance de la Inteligencia Artificial y más específicamente, del *Deep Learning*. Esta revolución ha permitido abordar una amplia gama de tareas. Se han desarrollado procedimientos más sofisticados para el reconocimiento, clasificación y tratamiento de imágenes médicas, lo que ha permitido mejorar significativamente el diagnóstico médico.

El atractivo principal del *Deep Learning* radica en su capacidad para adaptarse fácilmente a los datos, incluso ante pequeñas variaciones, sin necesidad de rediseñar por completo los algoritmos de aprendizaje. Esto ha marcado una diferencia fundamental con los métodos tradicionales. Es muy efectivo para tareas complejas como reconocimiento de imágenes y voz. [3]

En este contexto, las redes neuronales que están inspiradas en la estructura y función del cerebro. Han surgido como una herramienta poderosa para la extracción de características complejas en las imágenes, gracias a su estructura jerárquica compuesta por múltiples capas. Esta capacidad de modelado ha allanado el camino para aplicaciones médicas diversas, desde el diagnóstico por imagen hasta la predicción de resultados clínicos. [4]

Para comprender mejor este proyecto, brevemente se definirá a la sangre y sus componentes para luego explorar con más detalle cómo las nuevas arquitecturas, como los

Vision Transformer (*ViT*), están desafiando las convenciones establecidas por las redes neuronales convolucionales (*CNN*) y prometen ofrecer mejoras significativas en términos de eficiencia computacional y rendimiento predictivo.

En resumen, la convergencia de la Visión Artificial y el *Deep Learning* ha generado un cambio paradigmático en el campo de la salud, ofreciendo nuevas herramientas y enfoques que están transformando la manera en que se analizan y se interpretan las imágenes médicas, con el potencial de mejorar significativamente la atención médica y el diagnóstico clínico. [5]

II-A. Fundamentos de la Hematología y la importancia del análisis de la sangre periférica

II-A1. La sangre:

La sangre es un vehículo líquido de comunicación vital, entre los distintos tejidos del organismo; desempeña una variedad de funciones esenciales para mantener la salud y el equilibrio fisiológico. Está compuesta por una mezcla compleja de células suspendidas en un líquido llamado plasma, que contiene nutrientes, hormonas, electrolitos y otros compuestos necesarios para el funcionamiento adecuado del organismo. Entre otras funciones, destacan: [6]

- Distribución de nutrientes desde el intestino a los tejidos.
- Intercambio de gases: transporte de oxígeno desde los pulmones hasta los tejidos y de dióxido de carbono desde los tejidos hasta los pulmones.
- Transporte de productos de deshecho, resultantes del metabolismo celular, desde los lugares de producción hasta los de eliminación.
- Protección frente a microorganismos invasores.
- Protección frente a hemorragias.

II-A2. Componentes de la sangre: [7]

■ **Plasma:** El plasma representa alrededor del 55 % del fluido sanguíneo en los humanos. El plasma es principalmente agua que se absorbe de los alimentos ingeridos y los líquidos por los intestinos. Tiene 92 % agua, y el contenido del 8 % restante incluye:

- Glucosa
- Hormonas
- Proteínas
- Sales minerales
- Grasas
- Vitaminas

El 45 % restante de la sangre consiste principalmente en glóbulos rojos, glóbulos blancos y plaquetas. Cada uno de estos tiene un papel vital en mantener el funcionamiento efectivo de la sangre.

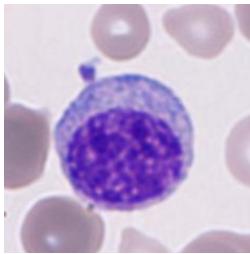


Figura 1: Imagen inmunoglobulina del *Dataset*

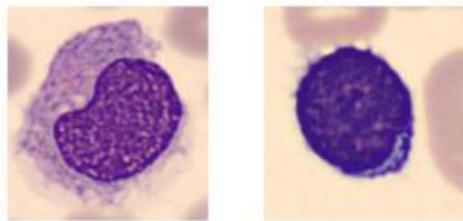


Figura 4: Monocito y linfocito del *Dataset*

- **Glóbulos rojos, o eritrocitos:** Tienen una forma ligeramente indentada y aplanaada en disco, que les permite viajar a través de los vasos sanguíneos de manera eficiente. Estas células son las más abundantes en la sangre y contienen hemoglobina, una proteína que transporta oxígeno desde los pulmones hacia los tejidos del cuerpo. La vida útil de un glóbulo rojo es de 4 meses, y el cuerpo los reemplaza regularmente. El cuerpo humano produce alrededor de 2 millones de glóbulos rojos por segundo.

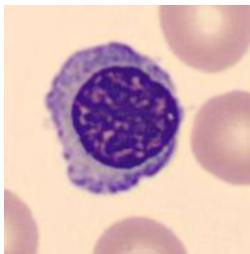


Figura 2: Imagen eritoblasto del *Dataset*

- **Glóbulos blancos, o leucocitos:** Son células del sistema inmunitario que protegen al cuerpo contra infecciones y enfermedades. representan menos del 1 % del contenido sanguíneo. El número de glóbulos blancos en un microlitro de sangre generalmente oscila entre 3,700–10,500. Los leucocitos se dividen en varios tipos, como los neutrófilos, linfocitos, monocitos, eosinófilos y basófilos, cada uno con funciones específicas en la respuesta inmunitaria.

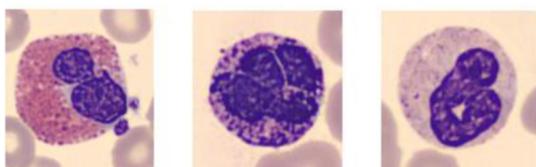


Figura 3: Eosinófilo, basófilo y neutrófilo del *Dataset*

- **Plaquetas, o trombocitos:** Son fragmentos celulares involucrados en la hemostasia, el proceso de coagulación sanguínea que detiene el sangrado cuando se produce una lesión en los vasos sanguíneos. Las plaquetas se adhieren al sitio de la lesión e interactúan con proteínas coagulantes para detener el sangrado. Debería haber entre 150,000 y 400,000 plaquetas por microlitro de sangre.

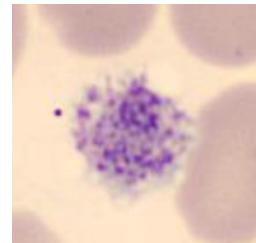


Figura 5: Imagen de plaquetas del *Dataset*

II-A3. *Importancia del análisis sanguíneo:*

Los análisis de sangre son una herramienta muy importante en el área médica, puesto que gracias a ellos el profesional de la salud tienen la posibilidad de diagnosticar diversas patologías; generalmente los pacientes llegan al laboratorio clínico para realizarse los estudios, sin saber el trabajo que existe detrás de ellos. Un diagnóstico médico depende en muchas ocasiones de análisis clínicos de calidad, en base a la precisión y confiabilidad de sus resultados, un especialista decide qué es lo mejor para tratar tal o cual enfermedad, así como lo más conveniente para el paciente, de acuerdo con sus características personales. [8]

Los análisis sanguíneos permiten una evaluación detallada de los componentes celulares de la sangre y pueden revelar desequilibrios en los niveles de glóbulos rojos, glóbulos blancos, plaquetas, así como la detección de anomalías morfológicas que pueden indicar enfermedades hematológicas, infecciones u otros trastornos.

El análisis microscópico del frotis sanguíneo teñido es una técnica específica utilizada para examinar las células sanguíneas; proporciona información acerca del número y forma de las células sanguíneas a través de una inspección visual con la ayuda del microscopio. Es útil como complemento de la biometría hemática que es de utilidad para establecer el diagnóstico de gran parte de las enfermedades hematológicas.

La sangre se obtiene por punción venosa; se coloca y extiende en un portaobjetos, de vidrio y teñida con colorantes especiales que resaltan las características morfológicas de las células, y son examinadas usando el microscopio. [9]

II-B. Modelos propuestos para la clasificación de células sanguíneas

II-B1. Redes Convolucionales (CNNs):

Las redes neuronales convolucionales (CNN) son redes complejas, forman parte esencial del aprendizaje profundo en el ámbito de la visión artificial. Estas redes han demostrado ser altamente eficaces en la resolución de problemas relacionados con la visión, desde la clasificación de imágenes hasta la interpretación de datos visuales y auditivos. [10]

En la mayoría de los casos, una red neuronal es un sistema que se adapta y cambia su estructura durante la fase de aprendizaje. Dicho aprendizaje, es una parte muy importante del proceso.

Las redes convolucionales combinan capas de convolución y *pooling* para extraer características importantes de los datos. La capa final, la *fully connected*, hace las predicciones. El *pooling* reduce la carga computacional y ayuda a evitar el sobreajuste. Sin embargo, en datos complejos, la reducción de dimensiones puede causar pérdida de información, llevando a errores como falsos positivos o negativos. [11]

Cuando las CNN son muy profundas, surgen problemas como la degradación o explosión del gradiente y la maldición de la dimensionalidad. El gradiente es esencial para el aprendizaje, pero puede volverse demasiado grande o pequeño, causando inestabilidad o deteniendo el aprendizaje. La maldición de la dimensionalidad ocurre cuando hay demasiadas características, lo que aumenta el tiempo de computación y puede hacer que la red no aprenda correctamente. [12]

II-B1a. Arquitectura de una Red Neuronal Convolutional:

La arquitectura de una CNN consta de capas convolucionales, capas de agrupamiento y capas totalmente conectadas. [13]

- Las capas convolucionales son responsables de detectar y extraer características importantes de una imagen mediante la aplicación de filtros convolucionales.
- Las capas de agrupamiento, también conocidas como capas de *pooling*, se utilizan para reducir la dimensionalidad de los mapas de características generados por las capas convolucionales, lo que ayuda a mantener la información relevante mientras se reduce el costo computacional.
- Las capas totalmente conectadas, al final de la red, se utilizan para aprender las relaciones complejas entre las

características identificadas por las capas convolucionales y las clases de salida, permitiendo así la clasificación adecuada de las imágenes.

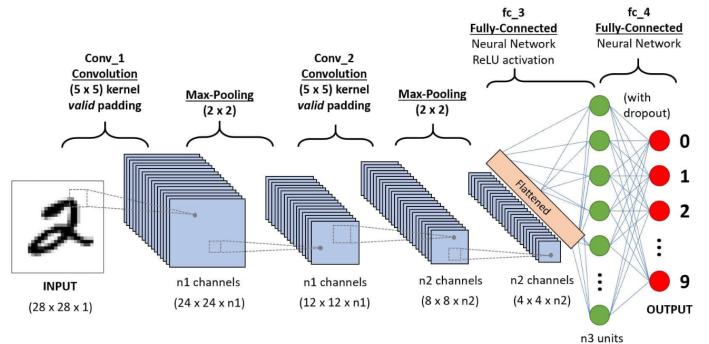


Figura 6: Arquitectura de CNN

II-B2. Vision Transformer (ViT) :

El modelo *Vision Transformer* (ViT) representa una innovación en el procesamiento de imágenes, basándose en los transformadores, que son componentes esenciales en el procesamiento del lenguaje natural (PNL). A diferencia de las convoluciones utilizadas en las redes neuronales convolucionales (CNN), los ViT dividen las imágenes en parches más pequeños y los procesan de manera individual mediante una arquitectura transformadora, lo que permite al modelo aprender y extraer información en diferentes niveles de abstracción, similar a su funcionamiento en tareas de PNL. [2]

Una característica distintiva de los ViT es su capacidad para manejar imágenes sin necesidad de preprocesamiento adicional, como las convoluciones. Aunque este enfoque puede parecer contraintuitivo al reducir la información en los parches de la imagen, ofrece beneficios significativos. La proyección lineal utilizada en los ViT para reducir la dimensionalidad de los parches de imagen resulta crucial en su arquitectura, lo que hace que el modelo sea más eficiente computacionalmente y accesible para investigadores y profesionales.

Al comparar los modelos de CNN y Vision Transformer, se observan diferencias notables en cuanto a tamaño del modelo, requisitos de memoria, precisión y rendimiento. Mientras que las CNN son conocidas por su tamaño compacto y eficiencia en el uso de la memoria, los Vision Transformer ofrecen una potente capacidad para capturar dependencias globales y comprensión contextual en imágenes, con un rendimiento mejorado en ciertas tareas. Sin embargo, los Vision Transformer tienden a tener modelos más grandes y requerimientos de memoria más altos, lo que puede limitar su practicidad en entornos con recursos limitados. La elección entre ambos modelos depende de las necesidades específicas de la tarea, considerando factores como los recursos disponibles, el tamaño del conjunto de datos y el equilibrio entre complejidad del modelo, precisión y rendimiento. Se anticipan nuevos avances en ambas arquitecturas, lo que permitirá tomar decisiones más

informadas basadas en necesidades y limitaciones específicas. [14]

II-B2a. Arquitectura del Transformador de Visión:

La estructura general del *Vision Transformer* sigue estos pasos simples: [15]

- Dividir la imagen en parches: La imagen se divide en pequeños fragmentos de tamaño fijo.
- Aplanar los parches de la imagen: Cada fragmento se aplana para formar una secuencia de datos.
- Crear incrustaciones lineales: Se crean representaciones lineales de menor dimensión a partir de estos fragmentos aplanados de la imagen.
- Incluir incrustaciones posicionales: Se añaden representaciones de posición a las incrustaciones lineales para que el modelo comprenda la disposición espacial de los parches.
- Alimentar la secuencia al codificador del *Transformer*: La secuencia resultante se introduce como entrada en un codificador *Transformer* avanzado.
- Entrenamiento previo del modelo: El modelo *ViT* se entrena inicialmente con etiquetas de imágenes en un conjunto de datos grande y supervisado.
- Ajuste posterior para la clasificación de imágenes: Se afinan los parámetros del modelo en conjuntos de datos específicos para la tarea de clasificación de imágenes.

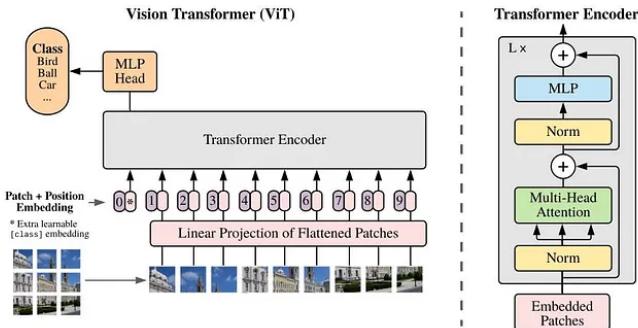


Figura 7: Ilustración *Transformer Encoder* fue inspirada por Vaswani et al. (2017)

Vision Transformers (ViT) es una arquitectura que utiliza mecanismos de autoatención para procesar imágenes. En lugar de analizar toda la imagen a la vez, *ViT* divide la imagen en pequeños fragmentos, los estudia individualmente y luego los combina para comprender la imagen completa. Utiliza una técnica llamada “autoatención” para examinar minuciosamente cada fragmento y entender cómo se relaciona con los demás. Además, *ViT* utiliza una capa especial para transformar esta información de manera significativa. [16]

ViT divide la imagen en pequeños “parches”, asignándoles números para facilitar su comprensión. Luego, procesa estos parches mediante bloques de procesamiento especializados para predecir qué contiene la imagen. Esta técnica permite que *ViT* proporcione una etiqueta que describe el contenido de la imagen.

Vision Transformers han demostrado ser extremadamente efectivos en varias tareas de visión artificial. Estos modelos utilizan mecanismos de autoatención de múltiples cabezales, lo que les permite manejar de manera flexible una secuencia de parches de imágenes para captar señales contextuales. Una ventaja clave de esta flexibilidad es su capacidad para enfrentar diversos problemas en imágenes reales, como occlusiones graves, cambios de dominio, permutaciones espaciales y perturbaciones tanto adversas como naturales.

Investigaciones recientes han explorado estas capacidades mediante un amplio conjunto de experimentos que incluyen tres familias de *ViT*, comparándolos con redes neuronales convolucionales (CNN) de alto rendimiento. Los resultados muestran que los *ViT* son notablemente resistentes a occlusiones, perturbaciones y cambios de dominio extremos. [17]

Esta robustez ante occlusiones no se debe a un sesgo hacia las texturas locales. De hecho, los *ViT* son mucho menos dependientes de las texturas en comparación con las CNN. Cuando se entranan adecuadamente para reconocer características basadas en formas, los *ViT* muestran una capacidad para identificar formas que es comparable al sistema visual humano, algo que no se había visto antes en la literatura. Además, esta capacidad permite una segmentación semántica precisa sin necesidad de supervisión a nivel de píxeles.

Otra ventaja es que las características obtenidas de un único modelo *ViT* pueden combinarse para formar un conjunto de características, lo que resulta en altas tasas de precisión en una variedad de conjuntos de datos de clasificación, tanto en paradigmas de aprendizaje tradicionales como de pocas oportunidades. Esta efectividad se debe a los campos receptivos flexibles y dinámicos que se logran a través de los mecanismos de autoatención. Nuestro código estará disponible públicamente para apoyar la investigación y el desarrollo en esta área. [18]

III. DISEÑO E IMPLEMENTACIÓN

III-A. Dataset

El estudio se basa en un conjunto de datos compuesto por 17,092 imágenes de células normales individuales, capturadas con el Cellavision DM96 en el Laboratorio Central del Hospital Clínico de Barcelona. Este repositorio ofrece varias ventajas, ya que las imágenes son de alta calidad y tienen las mismas dimensiones (363x360x3). Además, están organizadas en subdirectorios según los tipos celulares, lo que ahorra tiempo en el preprocesamiento necesario en muchos repositorios públicos.

Se podrían clasificar en dos enfoques: un primero enfoque en clasificar las imágenes en 8 grupos celulares principales, basados en los subdirectorios. El segundo enfoque utiliza 13 clases derivadas de los nombres de las imágenes. En este último, los grupos principales como neutrophil se subdividen en

band y segmented, mientras que ig se divide en promyelocyte, metamyelocyte y myelocyte. Finalmente, se opta por realizar el entrenamiento con las clases principales, es decir, 8 tipos celulares principales.

Para la preparación de los datos se definió como conjunto de test el 20 % de las imágenes. Las correspondientes etiquetas de clasificación se obtuvieron a partir de los nombres de cada uno de los archivos.

Además, se fijó el mismo tamaño para todas las imágenes con el fin de prevenir algún caso conflictivo en el que pudiera existir alguna ligera diferencia de dimensiones. [19]

III-B. Librerías y paquetes

- `import os`: Este módulo proporciona funciones para interactuar con el sistema operativo. Se utiliza para realizar operaciones relacionadas con archivos y directorios, como navegación de directorios, manipulación de rutas de archivos, etc.
- `import re`: Este módulo proporciona operaciones de coincidencia de expresiones regulares. Se utiliza para buscar patrones en cadenas de texto y realizar manipulaciones basadas en esos patrones.
- `import zipfile`: Este módulo proporciona clases para leer y escribir archivos ZIP. Se utiliza para trabajar con archivos comprimidos en formato ZIP, como extraer archivos de un archivo ZIP.
- `import torch`: Este es el módulo principal de PyTorch, un framework de aprendizaje profundo. Proporciona estructuras de datos y funciones para crear y entrenar redes neuronales.
- `import torch.nn as nn`: Este submódulo de PyTorch contiene las clases y funciones para definir y operar capas de redes neuronales. Se utiliza para construir modelos de redes neuronales con diferentes arquitecturas.
- `import torch.nn.functional as F`: Este submódulo proporciona funciones de activación y funciones de pérdida comunes utilizadas en el entrenamiento de redes neuronales.
- `from torch.utils.data import DataLoader, random_split`: Estas son clases de PyTorch para trabajar con conjuntos de datos y cargar datos en lotes durante el entrenamiento de redes neuronales. DataLoader se utiliza para cargar datos de un conjunto de datos, mientras que random_split se utiliza para dividir un conjunto de datos en conjuntos de entrenamiento y validación de manera aleatoria.
- `from torchvision import transforms`: Este módulo proporciona clases y funciones para realizar transformaciones de datos en imágenes, como cambiar de tamaño, recortar, rotar, etc. Se utiliza comúnmente para preprocessar imágenes antes de alimentarlas a modelos de redes neuronales.
- `from torchvision.datasets import ImageFolder`: Esta clase de torchvision se utiliza para crear un conjunto de datos de imágenes a partir de

una estructura de directorios donde cada subdirectorio representa una clase de imagen.

- `from torchvision import models`: Este módulo de torchvision proporciona implementaciones preentrenadas de modelos de redes neuronales para tareas de visión por computadora, como ResNet, VGG, etc.
- `import timm`: Este es el paquete que proporciona implementaciones de modelos de aprendizaje profundo preentrenados y herramientas relacionadas con la visión por computadora.
- `from sklearn.metrics import balanced_accuracy_score, cohen_kappa_score, confusion_matrix, ConfusionMatrixDisplay`: Estas son clases y funciones proporcionadas por scikit-learn, una biblioteca de aprendizaje automático en Python. Se utilizan para calcular métricas de evaluación de modelos de clasificación, como la precisión balanceada, el coeficiente kappa de Cohen y la matriz de confusión.
- `import matplotlib.pyplot as plt`: Este módulo se utiliza para crear gráficos y visualizaciones en Python.
- `import numpy as np`: Numpy es una biblioteca fundamental para la computación científica en Python. Proporciona soporte para matrices y operaciones matemáticas en ellas.
- `from PIL import Image, ImageOps`: El módulo PIL (Python Imaging Library) proporciona clases y funciones para abrir, manipular y guardar imágenes en diferentes formatos. Image se utiliza para operaciones básicas de imagen, mientras que ImageOps proporciona operaciones de procesamiento de imágenes como cambio de tamaño, rotación, etc.
- `import time`: Este módulo proporciona funciones para medir el tiempo de ejecución del código. Se utiliza para medir el tiempo que lleva ejecutar ciertas partes del código.

III-C. Diseño

III-C1. Estrategia de Implementación:

El objetivo principal de este proyecto es desarrollar un sistema capaz de clasificar imágenes en diferentes categorías con alta precisión. Para lograrlo, se emplea un enfoque basado en técnicas de aprendizaje profundo, específicamente utilizando el modelo *Vision Transformer (ViT)*. El modelo ViT es una arquitectura de redes neuronales que ha demostrado ser altamente efectiva en tareas de visión por computadora al tratar la entrada de imágenes como una secuencia de tokens. Esta técnica ha mostrado resultados prometedores en comparación con las arquitecturas convolucionales tradicionales, especialmente cuando se enfrenta a conjuntos de datos grandes y complejos.

III-C2. Preparación de datos:

Antes de entrenar un modelo de clasificación de imágenes, es crucial preparar los datos de manera adecuada. Esto implica cargar las imágenes desde un archivo comprimido, aplicar transformaciones para mejorar la calidad y la diversidad de los datos, y dividir el conjunto de datos en conjuntos de entrenamiento y validación. Esta división del conjunto de datos permite evaluar el rendimiento del modelo en datos que no ha visto durante el entrenamiento, proporcionando así una evaluación más realista de su capacidad para generalizar a nuevos datos.

III-C3. Definición del Modelo:

En este proyecto, se utiliza la arquitectura *Vision Transformer* (*ViT*) como modelo base para la clasificación de imágenes. El modelo *ViT* trata la entrada de imágenes como una secuencia de tokens, lo que le permite capturar relaciones de largo alcance entre los diferentes píxeles de la imagen. Esta arquitectura se elige por su capacidad para aprender representaciones de alto nivel de las imágenes y su eficacia en una variedad de tareas de visión por computadora.

Para adaptar el modelo *ViT* a nuestro problema específico de clasificación de imágenes, se reemplaza la capa de salida para que coincida con el número de clases en nuestro conjunto de datos. Además, se congela el resto de los parámetros del modelo preentrenado para evitar que se modifiquen durante el entrenamiento, lo que acelera el proceso y evita el sobreajuste.

III-C4. Entrenamiento del Modelo:

Durante el entrenamiento del modelo, se utiliza un optimizador y una función de pérdida para ajustar los parámetros con el objetivo de minimizar la pérdida en el conjunto de datos de entrenamiento.

Se realiza un seguimiento del progreso del entrenamiento mediante métricas como la pérdida, y se ajustan los hiperparámetros según sea necesario para mejorar el rendimiento del modelo. Además, se aprovecha la aceleración de GPU cuando está disponible para acelerar el proceso de entrenamiento y manejar conjuntos de datos más grandes de manera más efectiva, ya que se estaría trabajando en paralelo.

III-C5. Evaluación del Modelo:

Después de completar el entrenamiento del modelo, se evalúa su rendimiento en un conjunto de datos de validación separado. [20] Se calculan diversas métricas de evaluación, como la precisión, la precisión equilibrada y el coeficiente Kappa de Cohen, para evaluar la capacidad del modelo para realizar predicciones precisas en datos no vistos. Estas métricas proporcionan información sobre la capacidad del modelo para generalizar a nuevos datos y su capacidad para distinguir entre diferentes clases.

III-C6. Visualización de Resultados:

Para comprender mejor el rendimiento del modelo, se visualizan los resultados utilizando gráficos y representaciones visuales. Esto incluye gráficas de pérdida durante el entrenamiento para monitorear el progreso del modelo, matrices de confusión que muestran las predicciones del modelo frente a las etiquetas verdaderas, y visualizaciones individuales de muestras del conjunto de validación con sus etiquetas verdaderas y predicciones del modelo. Estas visualizaciones proporcionan información sobre los aciertos y errores del modelo, lo que ayuda a identificar áreas de mejora y a comprender su comportamiento en diferentes escenarios. En la siguiente imagen, muestra ejemplos de clasificación de células sanguíneas utilizando el modelo entrenado. Cada imagen está acompañada de las probabilidades predichas para cada clase por el modelo. Las etiquetas verdaderas y las predichas se indican en los títulos de las imágenes.

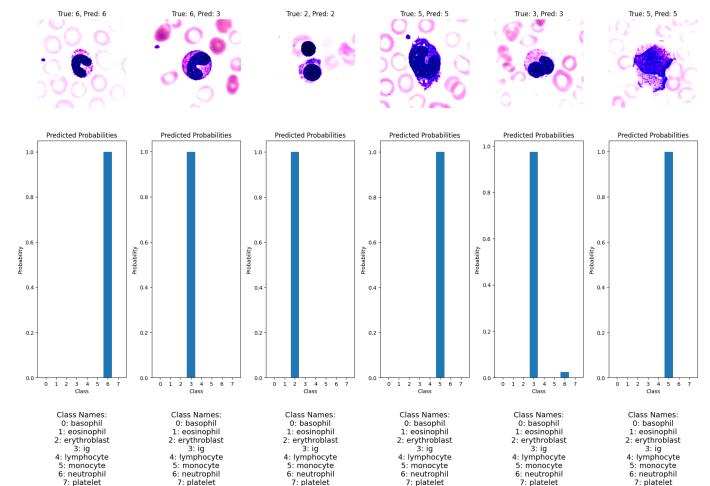


Figura 8: Clasificación de células sanguíneas con predicciones de probabilidad

III-D. Implementación

En una primera instancia, se toma por opción probar el modelo usando un tamaño de parche de 128 y luego de 64 usando Colab primero sin *GPU*, es decir, usando *CPU* lo cual habiendo pasado una cantidad mayor a 6 horas no se logra completar el proceso y se cancela automáticamente por uso excesivo de memoria *RAM* y por tiempo excedido para usar gratuitamente el servicio y configuraciones que nos facilita lo cual reducimos la cantidad de imágenes a procesar de 17,092 a 2,240 lo cual sigue con el mismo resultado de uso excesivo de memoria *RAM* y cancelación por tiempo de uso.

Para solucionar lo anterior se reduce el uso de la *RAM* disminuyendo el tamaño del parche a 32 pero aún con las 2,240 imágenes a procesar lo que se logra avanzar un poco

más y seguir con el proyecto pero también nos da el resultado anterior de uso excesivo de memoria *RAM* y tiempo excedido para usar gratuitamente el servicio.

Entonces teniendo los problemas anteriores ahora se utiliza la configuración con *GPU* para poder acelerar el proceso paralelizando el trabajo con el mismo tamaño de parche de 32 primero con 2,240 imágenes a procesar para primero validar si efectivamente se puede concluir con el proyecto de manera satisfactoria para lo cual el proceso se termina y demora aproximadamente 509.1546 segundos (8 minutos y 30 segundos aproximadamente) llegando a óptimos resultados llegando a un valor de 95.76 % en cuando a la precisión de validación, que mide la proporción de predicciones correctas sobre el total de predicciones.

Al final, teniendo resultados más óptimos y al completar el proceso, se decide realizar un nuevo proceso pero tomando las 17,092 imágenes dadas por el dataset original, lo cual nos da como resultado un tiempo de procesamiento de 3793.4880 segundos (1 hora y 3 minutos aproximadamente) llegando a óptimos resultados llegando a un valor de 96.84 % en cuando a la precisión de validación, que mide la proporción de predicciones correctas sobre el total de predicciones.

Después de analizar detenidamente esta implementación en cuanto al uso de *CPU* vs *GPU*, hemos identificado las siguientes observaciones:

■ Arquitectura y Diseño:

- Las *CPU* están diseñadas con un número menor de núcleos de procesamiento de propósito general, optimizados para manejar una variedad de tareas.
- Las *GPU* tienen una arquitectura altamente paralela con miles de núcleos más simples diseñados para procesar múltiples tareas simultáneamente.

■ Consumo de Energía:

- Las *CPU* suelen consumir menos energía en comparación con las *GPU*, lo que las hace más adecuadas para dispositivos móviles y sistemas con restricciones de energía.
- Las *GPU* pueden consumir mucha más energía debido a su arquitectura altamente paralela y al gran número de núcleos, lo que puede generar una mayor generación de calor y requerir sistemas de enfriamiento más avanzados.

■ Costo:

- Las *CPU* tienden a ser más caras por núcleo en comparación con las *GPU*.
- Las *GPU*, aunque pueden ser más caras en términos absolutos, ofrecen una mayor relación rendimiento-precio debido a su capacidad para manejar múltiples

tareas en paralelo.

■ Flexibilidad y Programación:

- Las *CPU* son más flexibles y pueden ejecutar una amplia gama de aplicaciones y algoritmos sin necesidad de optimizaciones especiales.
- Las *GPU* requieren programación específica para aprovechar al máximo su potencial, utilizando marcos como CUDA o OpenCL para desarrollar software optimizado para la arquitectura paralela de la *GPU*.

IV. RESULTADOS Y DISCUSIÓN

■ **Rendimiento del Modelo:** Los resultados del entrenamiento del modelo son bastante prometedores. Durante las 20 épocas de entrenamiento, se observa una disminución constante en la función de pérdida tanto en el conjunto de entrenamiento como en el de validación. Esto sugiere que el modelo está aprendiendo de manera efectiva y generalizando bien a datos no vistos. El coeficiente de *Kappa* de *Cohen* de validación también aumenta a lo largo del entrenamiento, indicando una mejor concordancia entre las predicciones y las etiquetas reales.

■ **Precisión y Métricas de Evaluación:** Se obtienen resultados bastante sólidos. La precisión de validación, que mide la proporción de predicciones correctas sobre el total de predicciones, se mantiene alta, alcanzando un valor máximo de alrededor del 96.84 %. Además, se observa que tanto la precisión equilibrada como el coeficiente *Kappa* de *Cohen* también alcanzan valores altos y estables, lo que indica que el modelo es capaz de clasificar de manera efectiva todas las clases de manera equitativa y consistente.

■ **Estabilidad y Generalización:** Aunque hubo algunas advertencias sobre el número de procesos trabajadores creados por el *DataLoader* y la incompatibilidad potencial con el código multihilo, lo que puede conducir a problemas de bloqueo y afectar la estabilidad del modelo, el proceso de entrenamiento fue relativamente estable y no se encontraron problemas significativos. Además, el modelo demostró una capacidad de generalización satisfactoria, ya que logró una precisión alta y consistente en el conjunto de validación.

■ **Tiempo de Ejecución:** El tiempo total de ejecución del código fue de aproximadamente 3793.4880 segundos (1 hora y 3 minutos aproximadamente). Aunque este tiempo puede variar según el hardware y otros factores, proporciona una indicación general del costo computacional asociado con el entrenamiento del modelo.

Para evaluar el rendimiento del modelo de clasificación en el conjunto de validación, se ha generado la matriz de confusión.

	basophil	eosinophil	erythroblast	ig	lymphocyte	monocyte	neutrophil	platelet
True label	236	0	0	3	1	1	0	0
predicted label	basophil	eosinophil	erythroblast	ig	lymphocyte	monocyte	neutrophil	platelet
basophil	236	0	0	3	1	1	0	0
eosinophil	0	647	0	1	0	3	1	0
erythroblast	0	1	291	1	2	1	2	1
ig	5	1	5	519	3	13	8	0
lymphocyte	0	0	0	1	237	2	1	0
monocyte	0	0	0	8	2	279	1	0
neutrophil	1	6	1	19	1	2	652	0
platelet	0	0	4	2	1	1	2	450

Figura 9: Matriz de confusión del rendimiento

Después de realizar pruebas con varias imágenes de internet, el modelo de clasificación de células sanguíneas demostró su eficacia al proporcionar resultados prometedores. A continuación se presentan algunos ejemplos destacados:

- Basófilo: Se logró una precisión del 94.86 %, lo que indica una clasificación precisa de esta célula sanguínea.

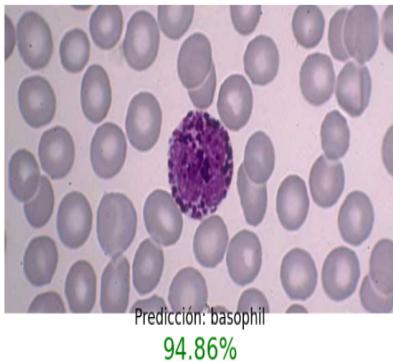


Figura 10: Imagen clasificada como Basófilo

- Eosinófilo: Se logró una precisión del 97.16 %, destacando una clasificación precisa de esta célula sanguínea.

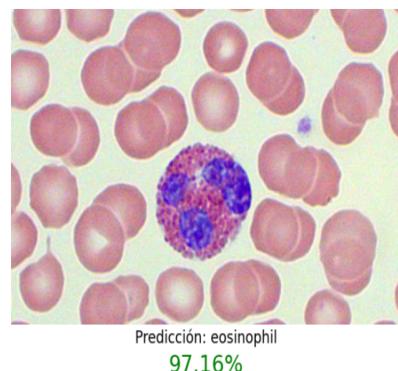


Figura 11: Imagen clasificada como Eosinófilo

- Eritroblasto: El modelo obtuvo una precisión del 100 % al identificar este tipo de célula, lo que refleja una clasificación perfecta.

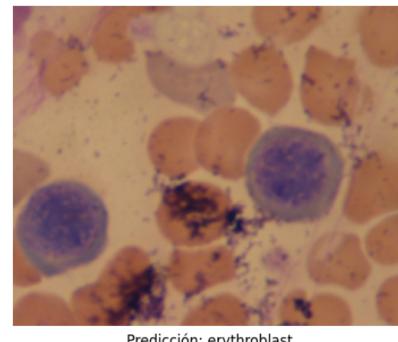


Figura 12: Imagen clasificada como Eritroblasto

- Linfocito: Se alcanzó una precisión del 95.04 % al identificar linfocitos, lo que demuestra una alta capacidad de clasificación para este tipo de células.

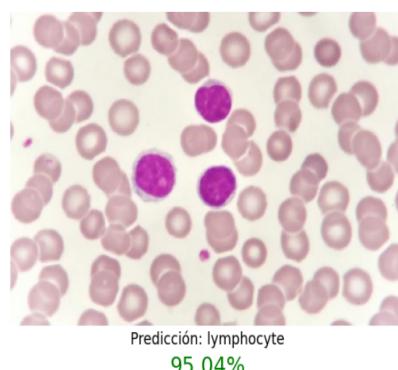
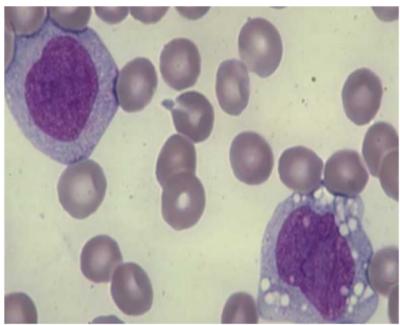


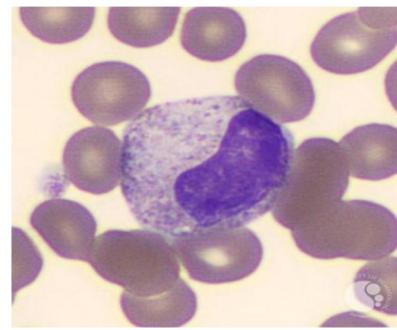
Figura 13: Imagen clasificada como Linfocito

- Monocito: El modelo predijo con una alta confianza al 95.60 % la presencia de monocitos en las imágenes procesadas, mostrando una identificación precisa de esta célula.



Predicción: monocyte
95.60%

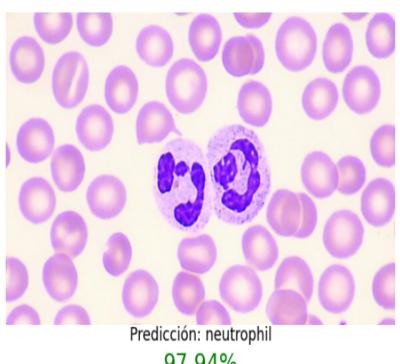
Figura 14: Imagen clasificada como Monocito



Predicción: ig
90.12%

Figura 17: Imagen clasificada como Ig

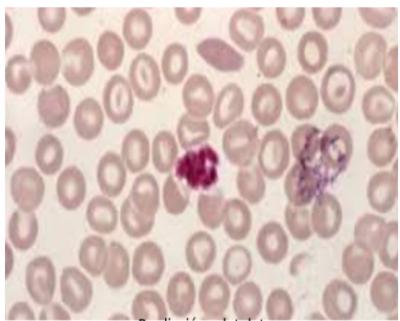
- Neutrófilo: Se obtuvo una precisión del 97.94 % al clasificar neutrófilos, lo que indica una clasificación muy precisa para este tipo de célula.



Predicción: neutrophil
97.94%

Figura 15: Imagen clasificada como Neutrófilo

- Plaquetas: La precisión alcanzada fue del 99.61 %, lo que demuestra una capacidad excepcional para identificar plaquetas en las imágenes.



Predicción: platelet
99.61%

Figura 16: Imagen clasificada como Plaquetas

- Ig Inmonoglobulina: Aunque con una precisión ligeramente menor, se logró un 90.12 % de precisión al identificar este tipo de célula, lo que aún indica una clasificación bastante precisa.

Estos resultados son indicativos del potencial del modelo para identificar y clasificar una variedad de células sanguíneas con alta precisión y confiabilidad.

Además de los resultados destacados anteriormente, se obtuvieron resultados similares y prometedores para otras imágenes de células sanguíneas. Si bien no se presentan en detalle, estos resultados respaldan la eficacia y precisión del modelo en la clasificación de una variedad de células sanguíneas, incluyendo las mencionadas anteriormente: basófilos, eritroblastos, linfocitos, monocitos, neutrófilos, plaquetas e Ig Inmonoglobulina. Se ha observado una clasificación precisa y confiable en estas células sanguíneas, lo que subraya la robustez y confiabilidad del modelo en su capacidad para identificar y clasificar células sanguíneas con precisión.

Tipo de Célula	Variante 1	Variante 2
Basófilos		
Eosinófilos		
Eritroblastos		
Linfocitos		

Cuadro I: Resultados adicionales de la clasificación de células sanguíneas (Parte 1)

Tipo de Célula	Variante 1	Variante 2
Monocitos		
Neutrófilos		
Plaquetas		

Cuadro II: Resultados adicionales de la clasificación de células sanguíneas (Parte 2)

V. CONCLUSIONES

- El entrenamiento del modelo mostró resultados prometedores. Durante las 20 épocas de entrenamiento, se observó una disminución constante en la función de pérdida tanto en el conjunto de entrenamiento como en el de validación, indicando que el modelo aprende eficazmente y generaliza bien a datos no vistos. El coeficiente de Kappa de Cohen también aumentó, mostrando una mejor concordancia entre las predicciones y las etiquetas reales.
- El modelo de clasificación de células sanguíneas mostró su eficacia con resultados prometedores en pruebas con varias imágenes de internet. Se logró una precisión alta en la clasificación de diferentes tipos de células sanguíneas, como basófilos, eosinófilos, eritroblastos, linfocitos, monocitos, neutrófilos y plaquetas, con precisiones que oscilan entre el 90.12 % y el 99 %.
- El tiempo total de ejecución del código fue de aproximadamente 3793.4880 segundos (aproximadamente 1 hora y 3 minutos). Aunque este tiempo puede variar según el hardware y otros factores, proporciona una indicación general del costo computacional asociado con el entrenamiento del modelo.
- Las pruebas realizadas demuestran la viabilidad y la eficacia del modelo de clasificación de células sanguíneas entrenado utilizando imágenes de internet. Este enfoque promete ofrecer importantes beneficios en el ámbito médico al simplificar y automatizar el proceso de detección y análisis de células sanguíneas, lo que resultaría en una mejora significativa de los diagnósticos médicos y potenciaría la investigación científica en este campo.

■ El entrenamiento y evaluación de modelos de clasificación de imágenes es un proceso complejo que requiere preparación cuidadosa, selección y adaptación del modelo, y evaluación exhaustiva. Un enfoque sistemático y herramientas adecuadas permiten crear modelos efectivos para diversos escenarios en el mundo real. El uso de *Vision Transformer (ViT)* en este proyecto destaca su eficacia y versatilidad en tareas complejas de visión por computadora, mostrando su potencial en este campo en constante evolución.

■ Para tareas que requieren flexibilidad y la capacidad de manejar una amplia variedad de aplicaciones, las *CPU* son la opción más adecuada debido a su diseño versátil y su capacidad para ejecutar diversos algoritmos. Por otro lado, las *GPU* son ideales para aplicaciones que se benefician de un procesamiento altamente paralelo, como el aprendizaje profundo y la simulación de fluidos, gracias a su arquitectura masivamente paralela.

■ Las *CPU* son más fáciles de programar y pueden ejecutar una amplia gama de aplicaciones sin necesidad de optimizaciones especiales. En contraste, las *GPU* requieren programación específica para aprovechar al máximo su arquitectura paralela, lo que puede implicar un aprendizaje adicional y el uso de marcos de programación como *CUDA* u *OpenCL*.

■ Aunque la falta de un equipo con gran potencia de procesamiento puede ser una limitación, los usuarios pueden explorar técnicas de Deep Learning mediante plataformas como *Colab* o *Google Cloud*, utilizadas en este proyecto. Sin embargo, disponer de un ordenador propio con un entorno de ejecución estable presenta ventajas significativas, como la eliminación de restricciones de tiempo y la retención de los datos generados durante el entrenamiento, evitando su pérdida al cerrar la sesión. Esto, no obstante, implica un mayor presupuesto.

■ La infraestructura de hardware juega un papel crucial en el rendimiento y la eficiencia del modelo de clasificación. Contar con *GPU* de última generación puede acelerar significativamente el proceso de entrenamiento y permitir la implementación de modelos más complejos y precisos. Por tanto, la inversión en hardware adecuado es fundamental para aprovechar al máximo las capacidades del modelo *Vision Transformer (ViT)* y otros algoritmos avanzados de aprendizaje profundo.

REFERENCIAS

- [1] R. C. Joshi, S. Yadav, M. K. Dutta1, and C. M. Travieso-Gonzalez, “An efficient convolutional neural network to detect and count blood cells,” *Uniciencia*, 2022.
- [2] O. Katar and O. Yildirim, “An explainable vision transformer model based white blood cells classification and localization,” *MDPI*, 2023.

- [3] R. Barrere, L. Matas, J. Sokil, and L. Trama, *Dossier: Inteligencia Artificial*. Observatorio Iberoamericano de la Ciencia, la Tecnología y la Sociedad.
- [4] X. B. Olabe, *Redes neuronales artificiales y sus aplicaciones*. Escuela Superior de Ingeniería de Bilbao, EHU.
- [5] R. Aggarwal, V. Sounderahajah, G. Martín, D. S. Ting, A. Karthikesalingam, D. Rey, H. Ashrafi, and A. Darzi. (2021) Precisión diagnóstica del aprendizaje profundo en imágenes médicas: una revisión sistemática y un metanálisis. [Online]. Available: <https://www.nature.com/articles/s41746-021-00438-z>
- [6] J. R. Palacios, "Sistema inmune y la sangre," *Colegio Oficial Enfermeres, Barcelona*.
- [7] A. Felman. (2024) How does blood work, and what problems can occur? [Online]. Available: <https://www.medicalnewstoday.com/articles/196001>
- [8] E. de Marketing. (2018) La importancia de los análisis de sangre. [Online]. Available: <https://cedimmont.com/blog/la-importancia-de-los-analisis-de-sangre/>
- [9] J. C. Pérez and D. G. Almaguer, *Hematología. La sangre y sus enfermedades. Cap 55 Frotis de la sangre periférica en las enfermedades más frecuentes*. McGraw-Hill Interamericana Editores, S. A, 2015.
- [10] I. A. J. Trujillo, J. P. Z. de Paz, O. P. Sandoval, and F. A. C. Velásquez, "Calibración de cámara multiespectral utilizando redes neuronales convolucionales," *Universidad Politécnica de Querétaro. Computación y Sistemas*, 2023.
- [11] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 548–558, 2021. [Online]. Available: <https://api.semanticscholar.org/CorpusID:232035922>
- [12] L. Moreno Diaz Alejo, "Análisis comparativo de arquitecturas de redes neuronales para la clasificación de imágenes," *Universidad Internacional de La Rioja (UNIR)*, 2020.
- [13] Fineproxy.org. Redes neuronales convolucionales (cnn). [Online]. Available: <https://fineproxy.org/es/wiki/convolutional-neural-networks-cnn/>
- [14] F. Rustamy. (2023) Vision transformers vs. convolutional neural networks. [Online]. Available: <https://medium.com/@faheemrustamy/vision-transformers-vs-convolutional-neural-networks-5fe8f9e18efc>
- [15] G. Boesch. (2024) Vision transformers (vit) in image recognition. [Online]. Available: <https://viso.ai/deep-learning/vision-transformer-vit/>
- [16] C. P. Lee, K. M. Lim, Y. X. Song, and A. Alqahtani, "Plant-cnn-vit: Plant classification with ensemble of convolutional neural networks and vision transformer," *Plants*, vol. 12, 2023. [Online]. Available: <https://api.semanticscholar.org/CorpusID:259911794>
- [17] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, F. E. H. Tay, J. Feng, and S. Yan, "Tokens-to-token vit: Training vision transformers from scratch on imagenet," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 538–547, 2021. [Online]. Available: <https://api.semanticscholar.org/CorpusID:231719476>
- [18] M. Naseer, K. Ranasinghe, S. H. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Intriguing properties of vision transformers," in *Neural Information Processing Systems*, 2021. [Online]. Available: <https://api.semanticscholar.org/CorpusID:235125781>
- [19] B. R. Maciejewska, "Exploración de vision transformer para la clasificación de células normales de sangre periférica," *Universitat Oberta de Catalunya*, 2021.
- [20] Y. Li, S. Xu, B. Zhang, X. Cao, P. Gao, and G. Guo, "Q-vit: Accurate and fully quantized low-bit vision transformer," *ArXiv*, vol. abs/2210.06707, 2022. [Online]. Available: <https://api.semanticscholar.org/CorpusID:252873138>