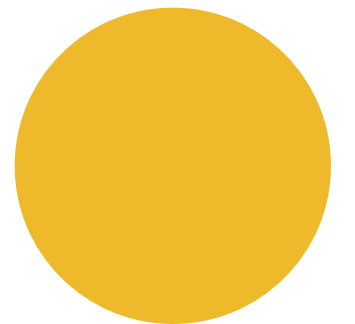


Detección de anemia en micrografía sanguínea por medio de análisis multimodal

Proyectos VI

- Andrea de los Angeles Mena Uicab
- Gabriela Jerez Gomez
- Ana Cristina Castro Canchola



INTRODUCCIÓN

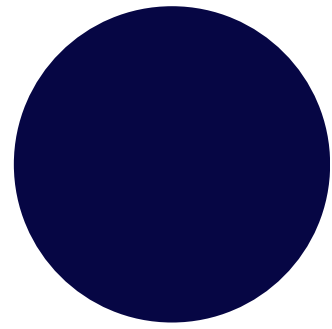
En México, la Encuesta Nacional de Salud y Nutrición (Ensanut), reporta que la prevalencia de anemia alcanza 15.8 % en mujeres de 20 a 49 años, además de afectar a 10.1 % de los adolescentes y 10.3 % de los adultos mayores [1].

Su diagnóstico suele basarse en análisis de laboratorio y en la observación de frotis sanguíneos mediante microscopía [2].

Ante esta problemática, surge la necesidad de desarrollar un modelo multimodal para la detección de anemia con un marco de aprendizaje profundo de dos etapas para la segmentación y clasificación de imágenes de glóbulos rojos (GR).

[1] Mejía-Rodríguez F, Mundo-Rosas V, García-Guerra A, Mauricio-López ER, Shamah-Levy T, Villalpando S, et al. Prevalence of anemia in the Mexican population: the analysis of Ensanut Continua 2022. *Salud Publica Mex* 2023;65. <https://doi.org/10.21149/14771>.

[2] Lee, S. Y., Chen, C. M. E., Lim, E. Y. P., Shen, L., Sathe, A., Singh, A., Sauer, J., Taghipour, K., & Yip, C. Y. C. (2021). Image analysis using machine learning for automated detection of hemoglobin H inclusions in blood smears. *Journal of Pathology Informatics*, 12, 18. https://doi.org/10.4103/jpi.jpi_110_20



ANTECEDENTES

Referencia	Dataset	Numero de imagenes/variables cuantitativas	Algoritmos	Accuracy	Conclusión
1	HAM10000 dataset	El conjunto de datos final comprende 10,015 imágenes dermatoscópicas. También contiene metadatos para cada paciente, incluyendo datos clínicos.	ClinicalMLP, DermiResNet, Bloque de fusión, Clasificación	Metadatos: 77%. Imágenes: 92%. Multimodal: Cross-attention y Hadamard product alcanzaron 98.86% y 98.85%, respectivamente.	La integración multimodal mejora significativamente la exactitud del diagnóstico de lesiones cutáneas

OBJETIVO GENERAL

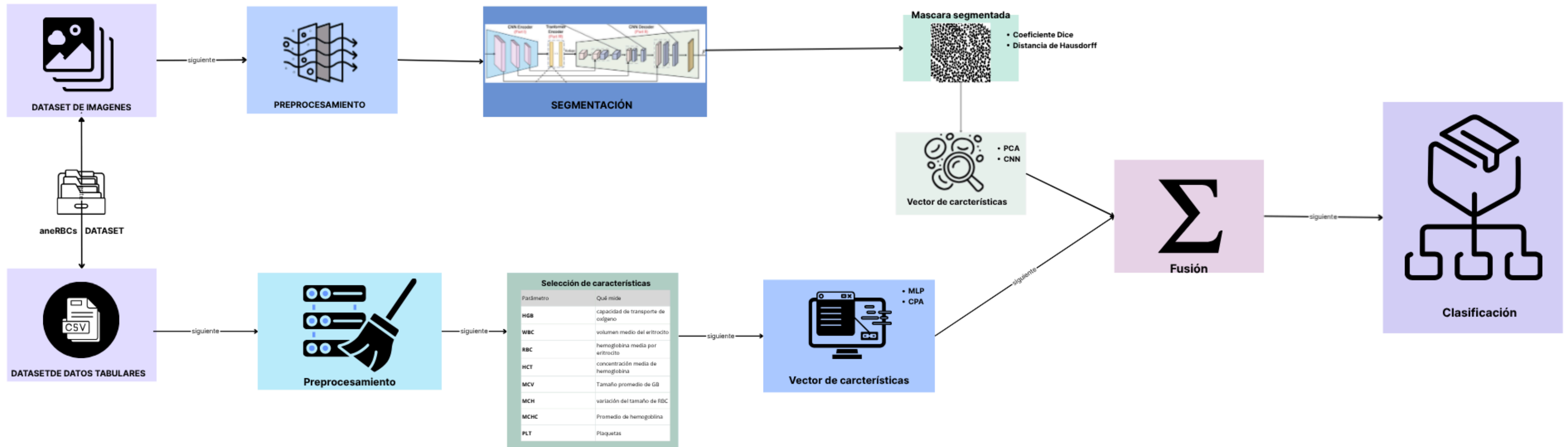
Desarrollar un sistema de detección de anemia basado en aprendizaje profundo utilizando un enfoque multimodal que combine datos hematológicos e imágenes de microscopía de frotis sanguíneo obtenidas de un mismo dataset.

OBJETIVO ESPECIFICOS

- *Analizar datasets que se conformen por datos hematológicos e imágenes de frotis sanguíneo principalmente orientados a anemia.*
- *Entrenar un modelo U-Net para segmentar eritrocitos en imágenes de frotis sanguíneo del dataset AneRBC*
- *Identificar la información hematológica relevante mediante la selección de características a partir de datos clínicos tabulares*
- *Implementar una fusión intermedia de los vectores de imágenes y datos cuantitativos para la clasificación final de anemia.*



Diagrama del Proyecto



Análisis de datasets

REFERENCIA	Identificación	Tipo de datos	Volumen	Accesibilidad
1	Anemia Dataset	<ul style="list-style-type: none">• Variables Hematológicas: <i>Hemoglobina, MCHC, MCV, MCH</i>• Demografía de Pacientes: sexo• Etiqueta de resultado	4 Variables Hematológicas: <ul style="list-style-type: none">• 1421 muestras	Acceso Publico
2	Pediatric Anemia Dataset: Hematological Indicators and Diagnostic Classification	<ul style="list-style-type: none">• Pacientes Pediátricos• Variable de resultado• Demografía de Pacientes: <i>Sexo ,edad.</i>• Parámetros hematológicos: clave: <i>Hb, RBC, PCV, MCV, MCH, MCHC</i>	1000 muestras	Acceso Publico

1. <https://www.kaggle.com/datasets/biswaranjanrao/anemia-dataset>

2. Mojumdar, M. U., Sarker, D., Assaduzzaman, M., Rahman, M. M., Sajeeb, M. A. H., Shadikul, B. . M., Siddiquee, S. M. T., & Chakraborty, N. R. (2024). Pediatric Anemia Dataset: Hematological Indicators and Diagnostic Classification [Conjunto de datos]. En Data Archiving and Networked Services (DANS). <https://doi.org/10.17632/y7v7ff3wpj.1>

Análisis de datasets

REFERENCIA	Identificación	Tipo de datos	Volumen	Accesibilidad
3	AneRBC-1	<ul style="list-style-type: none">• Imágenes del frotis sanguíneo• Variables Hematológicas: HCT,MVC,MCH,MCHC,RDW,PLT, MPV• Variable de resultado	<ul style="list-style-type: none">• 1000 imágenes originales• 1000 mascarar binarias• 1000 mascarar RGB• 17 Variables Hematológicas: 1000 muestras	Acceso Publico
4	CP-AnemiC	<ul style="list-style-type: none">• Pacientes pediátricos (6 a 59 meses)• Imágenes de conjuntiva• Metainformación: Hb , sexo, edad, observaciones y lugar de recolección	710 imágenes: <ul style="list-style-type: none">• anémicos(424)• no anémicos(286)	Acceso Publico

Resultados | Preprocesamiento

id	HCT	MCV	MCH	MCHC	RDW	PLT	MPV	%NEUT	%LYMP	%MONO	%EOS	%BASO	#NEUT	#LYMP	#MONO	#EOS	#BASO
001_h	41.9	90.1	29	32.2	12.8	331	11	59	34.4	5	1.1	0.5	4.36	2.54	0.08	0.37	0.04
002_h	34.4	78	26.3	33.7	12.5	251	11.4	73.7	14.9	7	4	0.4	8.06	1.63	0.76	0.44	
003_h	50.4	85.6	30.1	35.1	12.5	180	13.2	47.5	35.3	14.5	2.2	0.5	3.05	2.26	0.93	0.14	0.03
004_h	41.9	90.1	29	32.2	12.8	331	11	59	34.4	5	1.1	0.5	4.36	2.54	0.37	0.08	
005_h	39.5	98.8	34	34.4	12.9	102	13.1	53	37.1	7.4	1.9	0.6	3.61	2.52	0.5		0.04
006_h	51.8	84.2	27	32	13.2	271	11.5	60.9	27.4	8.3	3	0.4	7.27	3.27	0.99	0.36	0.05
007_h	41.6	88.5	31.3	35.3	14.9	496	10.5	35.3	47.5	13.7	3.1	0.4	5.12		1.99		0.06
008_h	34.8	83.7	31.3	37.4	12.2	40	11.5	89.7	6.2	4.1	0	0	1.31	0.09	3	0.02	
009_h	45	83.2	29.6	35.6	13.2	247	11.9	41.2	47.9	5	5.6	0.3	3.75	4.37	0.46	0.51	0.03
010_h	37.6	90.4	28.8	31.9	13.2	278	11.8	64.7	23.7	6.9	4.4	0.3	5.55	2.03	0.59	0.38	0.03
011_h	43.2	78.7	25.5	32.4	14.6	345	11.3	61.1	24.5	12.6	1.4	0.4	4.93	1.98	1.02	0.11	0.03
012_h	47.5	76.1	24.5	32.2		295	9.8										
013_h	36	81.1	25	30.8	14.3	369	11.3	60.8	31.2	7.1	0.4	0.5	5.04	2.59	0.59	0.03	0.04
014_h	44.1	82	27.5	33.6		302	11.4										
015_h	23.7	81.7	27.9	34.2	19.5	48											
016_h	35.5	86.4	30.7	35.5	12.9	201	12.9	52.3	36.6	8.5	2	0.6	2.65	1.85	0.43	0.1	0.03
017_h	39.9	84.4	30.2	35.8	13.2	252	10.5	53	33.7	6.1	6.5	0.7	3.2	2.03			0.04
018_h	36	74.7	25.1	33.6	14.8	359	8.7	36.7					3.6	5.38		0.02	0.03
019_h	35.4	74.1	24.5	33.1	15.8	394	9	14.8	77.7	4.4	2.9	0.2	1.85	9.72	0.55	0.36	0.03
020_h	48.7	84.5	29.3	34.7	12.5	191	11.6	59.7	27.9	8.7	3.3	0.4	3.3	1.54	0.48	0.18	0.02
021_h	41.2	89	30.2	34	13.7	287	10.5	45.7	42.2	7.2	4.6	0.3	3.18	0.0293	0.5	0.32	0.02
022_h	36.1	82.4	25.1	30.5		15											
023_h	36.4	86.7	29.8	34.3		221	13.1										
024_h	39.3	80.9	26.7	33.1	16.4	207	12.7	5.2	67.8	26.2	0.3	0.5	6.14	78.93	30.52	0.34	0.55

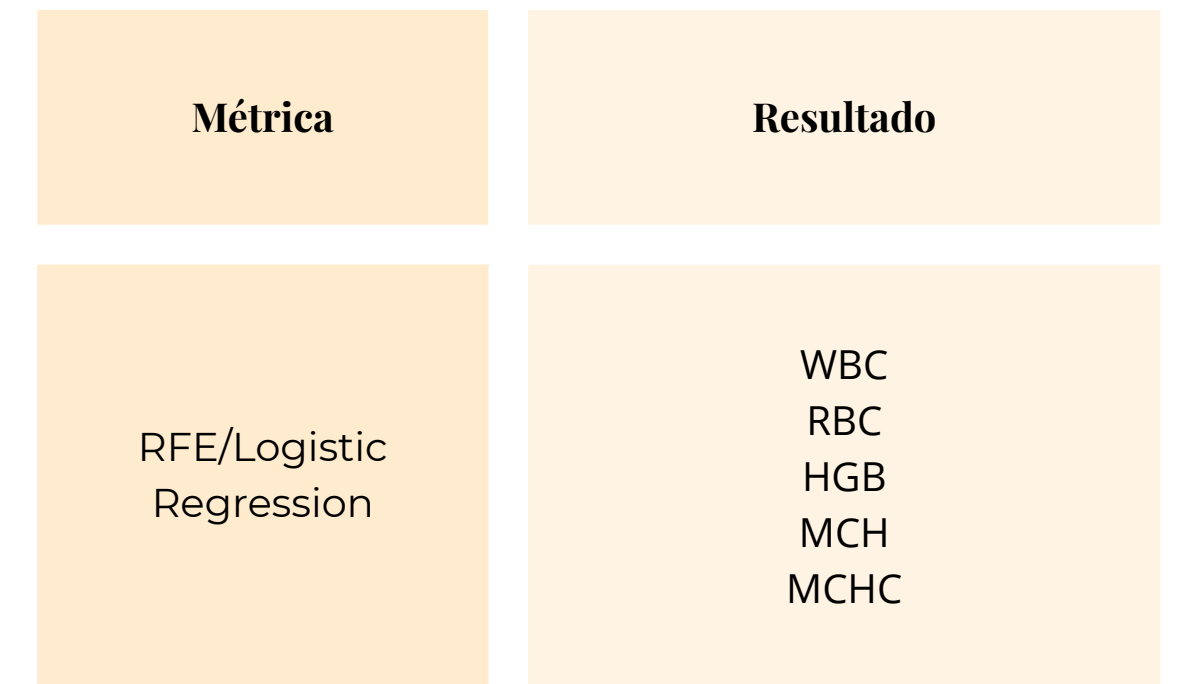
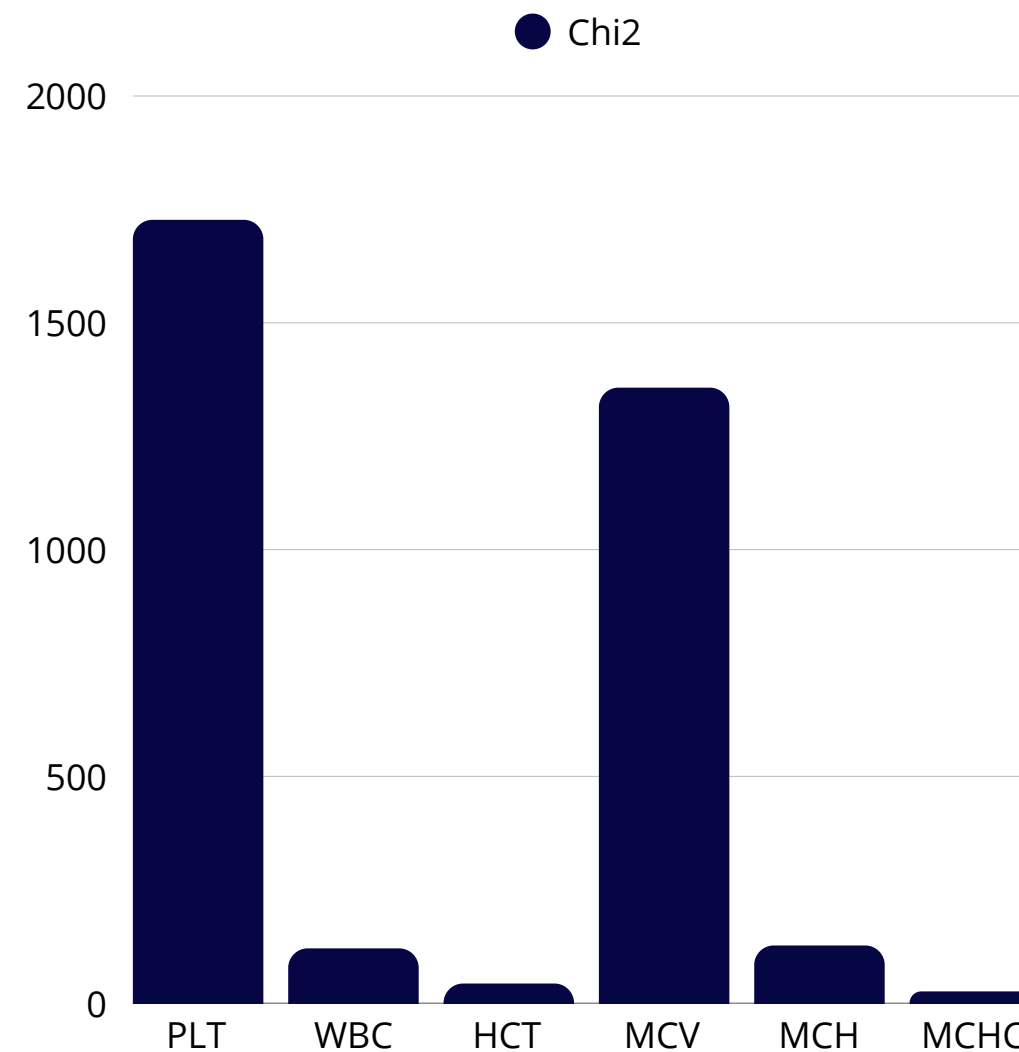
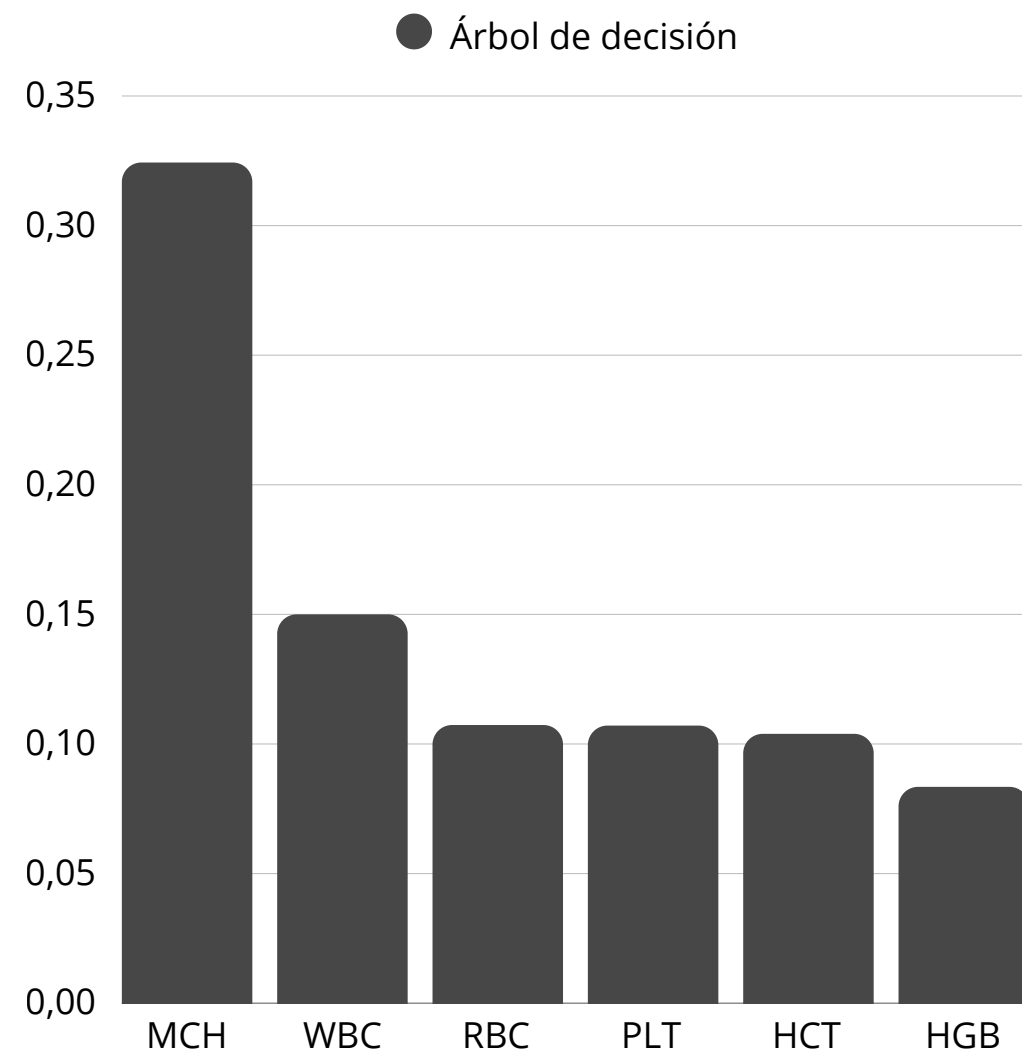
Los datos hematológicos de todas las muestras fueron consolidados en un único archivo Excel, se identificaron 17 variables originales de las cuales se eliminaron las que presentaban mayor cantidad de muestras sin valor. Se eliminaron las mismas muestras en las imágenes de frotis sanguíneo.

Resultados | Preprocesamiento

Al eliminar dichas variables nos quedamos con 8 características, es decir con 485 muestras anémicas y la misma cantidad para saludables, estos datos se consolidaron en un archivo nuevo de Excel.

id	WBC	RBC	HGB	HCT	MCV	MCH	MCHC	PLT
001_h	7.39	4.65	13.5	41.9	90.1	29	32.2	331
002_h	10.93	4.41	11.6	34.4	78	26.3	33.7	251
003_h	6.41	5.89	17.7	50.4	85.6	30.1	35.1	180
004_h	7.39	4.65	13.5	41.9	90.1	29	32.2	331
005_h	6.8	4	13.6	39.5	98.8	34	34.4	102
006_h	11.94	6.15	16.6	51.8	84.2	27	32	271
007_h	14.51	4.7	14.7	41.6	88.5	31.3	35.3	496
008_h	1.46	4.16	13	34.8	83.7	31.3	37.4	40
009_h	9.12	5.41	16	45	83.2	29.6	35.6	247
010_h	8.58	4.16	12	37.6	90.4	28.8	31.9	278
011_h	8.07	5.49	14	43.2	78.7	25.5	32.4	345
012_h	7.73	6.24	15.3	47.5	76.1	24.5	32.2	295
013_h	4	8.29	11.1	36	81.1	25	30.8	369
014_h	5.94	5.38	14.8	44.1	82	27.5	33.6	302
015_h	0.61	2.9	8.1	23.7	81.7	27.9	34.2	48
016_h	5.06	4.11	12.6	35.5	86.4	30.7	35.5	201
017_h	6.03	4.73	14.3	39.9	84.4	30.2	35.8	252
018_h	9.8	4.82	12.1	36	74.7	25.1	33.6	359
019_h	12.51	4.78	11.7	35.4	74.1	24.5	33.1	394

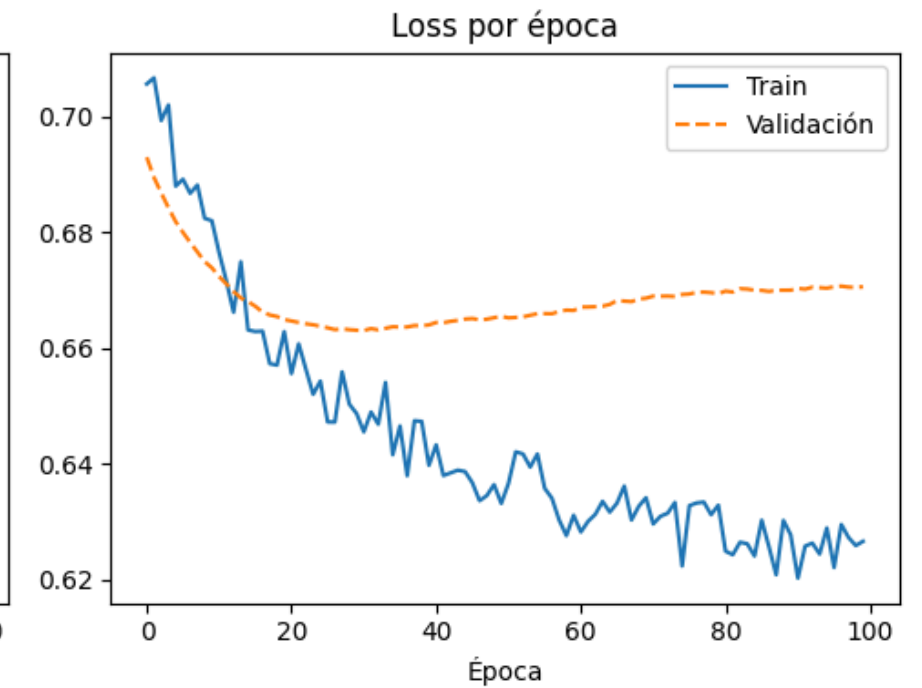
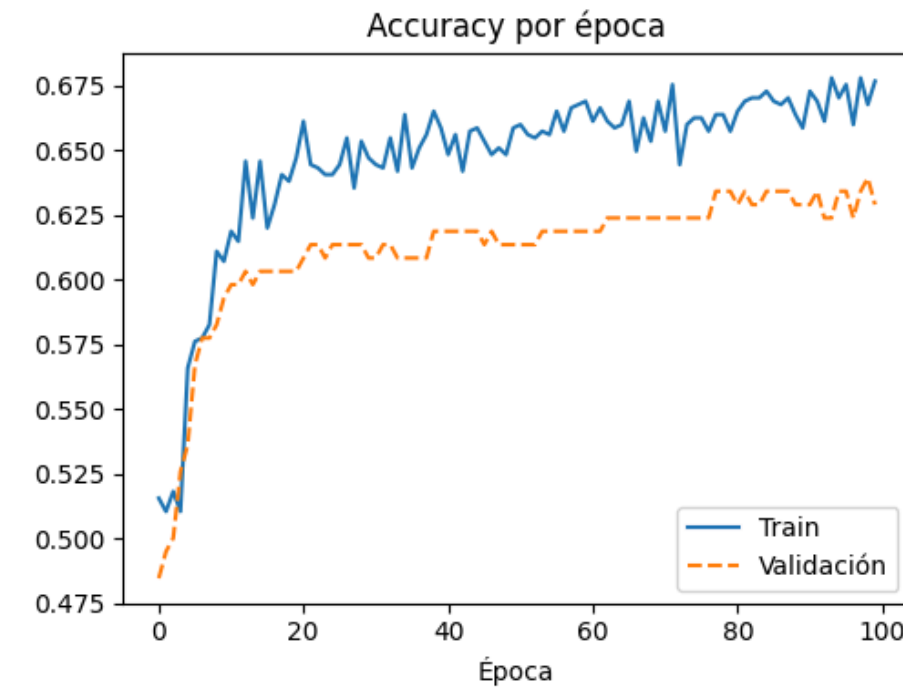
Simulación | Preprocesamiento



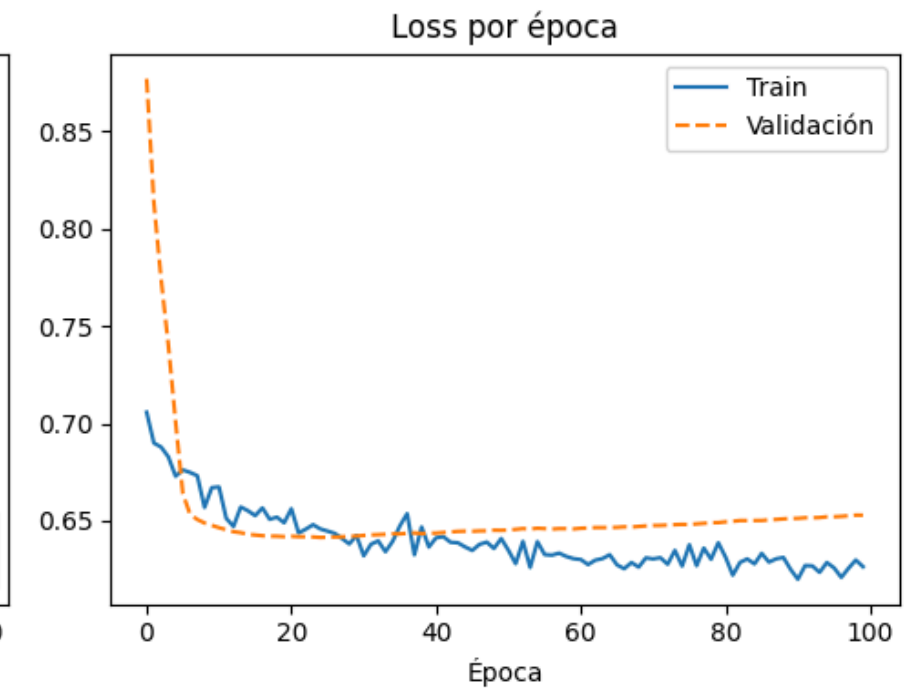
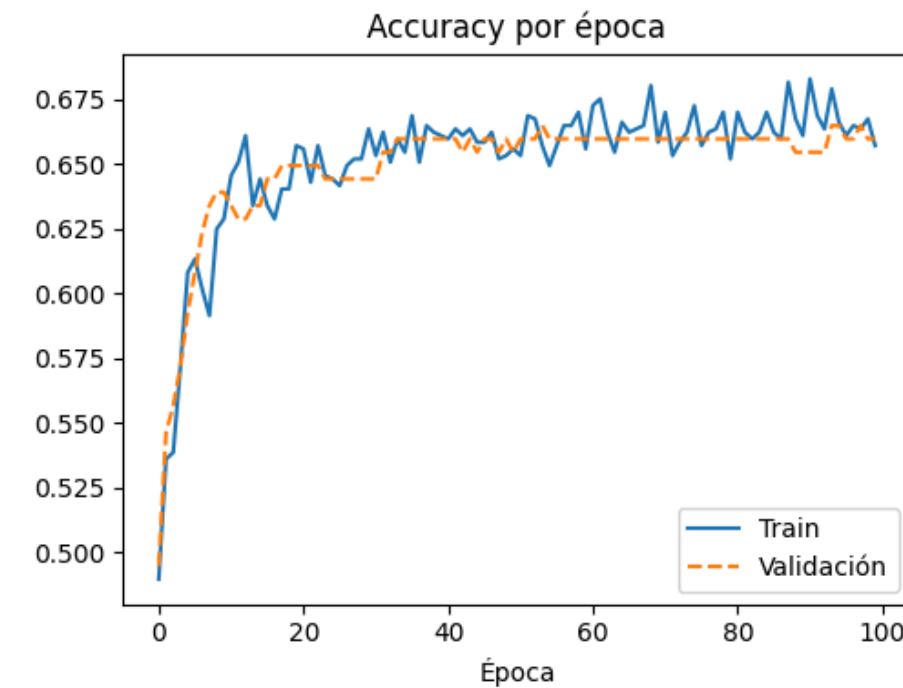
Se realizó una selección de datos mediante 3 algoritmos: Árbol de decisión, Chi2 y Eliminación Recursiva de Características con Regresión Logística.

Simulación | Preprocesamiento

Utilizando las 6 variables del hemograma como entrada el modelo alcanzó un accuracy de 62.89% en validación y con una pérdida de 0.6706 respectivamente.



Utilizando las 8 variables del hemograma como entrada el modelo alcanzó un accuracy de 65.93% en validación y con una pérdida de 0.6573 respectivamente.



Model: "functional"

Layer (type)	Output Shape	Param #	Connected to
input_layer (InputLayer)	(None, 300, 300, 3)	0	-
conv2d (Conv2D)	(None, 300, 300, 32)	896	input_layer[0][0]
max_pooling2d (MaxPooling2D)	(None, 150, 150, 32)	0	conv2d[0][0]
conv2d_1 (Conv2D)	(None, 150, 150, 64)	18,496	max_pooling2d[0]_
max_pooling2d_1 (MaxPooling2D)	(None, 75, 75, 64)	0	conv2d_1[0][0]
conv2d_2 (Conv2D)	(None, 75, 75, 128)	73,856	max_pooling2d_1[
conv2d_transpose (Conv2DTranspose)	(None, 150, 150, 64)	32,832	conv2d_2[0][0]
concatenate (Concatenate)	(None, 150, 150, 128)	0	conv2d_transpose_
conv2d_3 (Conv2D)	(None, 150, 150, 64)	73,792	concatenate[0][0]
conv2d_transpose_1 (Conv2DTranspose)	(None, 300, 300, 32)	8,224	conv2d_3[0][0]
concatenate_1 (Concatenate)	(None, 300, 300, 64)	0	conv2d_transpose_
conv2d_4 (Conv2D)	(None, 300, 300, 1)	65	concatenate_1[0]_

Total params: 208,161 (813.13 KB)
Trainable params: 208,161 (813.13 KB)
Non-trainable params: 0 (0.00 B)

Resultados

Segmentación de imágenes de frotis sanguíneo

Se adoptó la arquitectura UNet propuesta por Shahzad et al. (2024) en el artículo AneRBC dataset: a benchmark dataset for computer-aided anemia diagnosis using RBC images. El modelo fue entrenado durante 50 épocas con un batch size de 8.

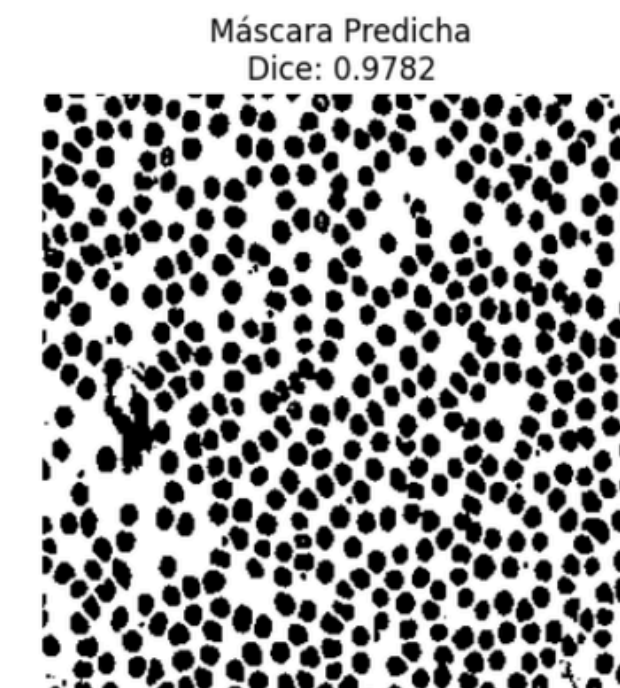
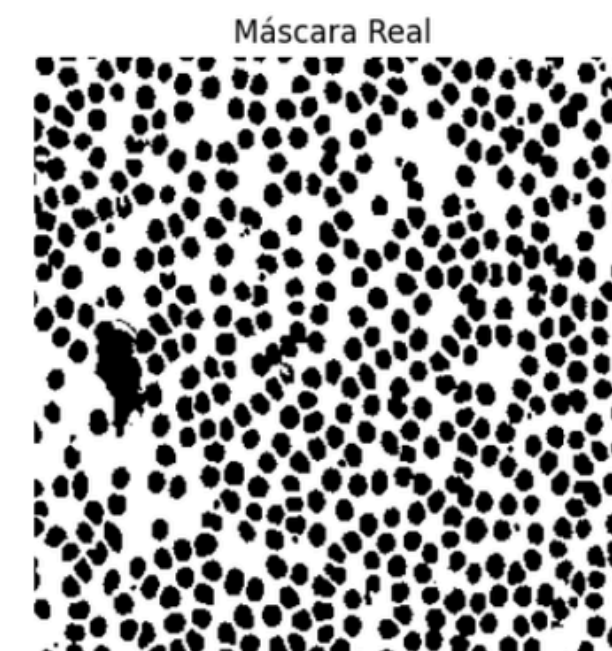
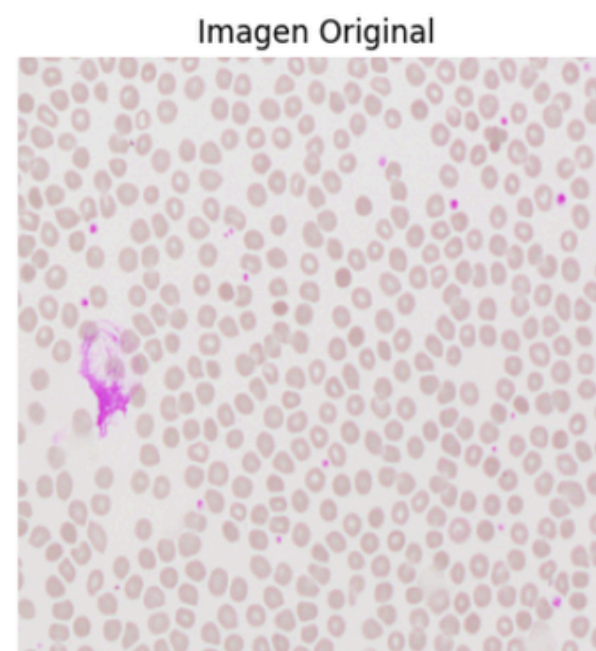
Simulación

Segmentación de imágenes de frotis sanguíneo

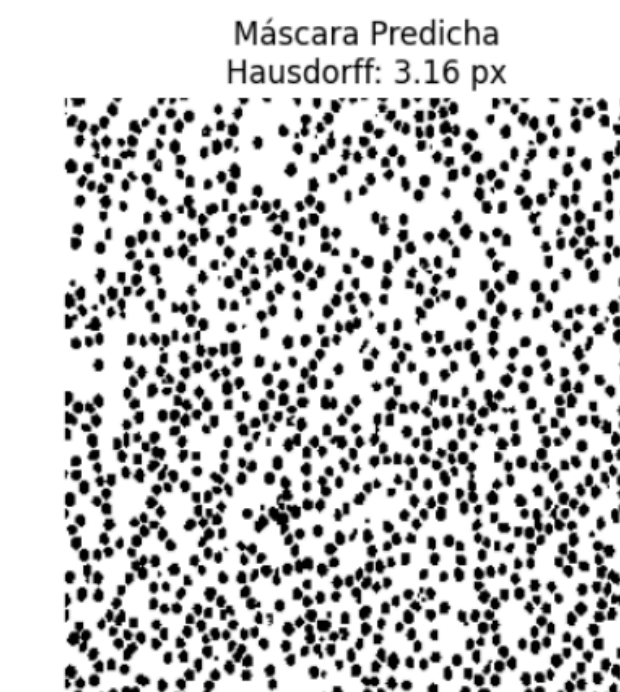
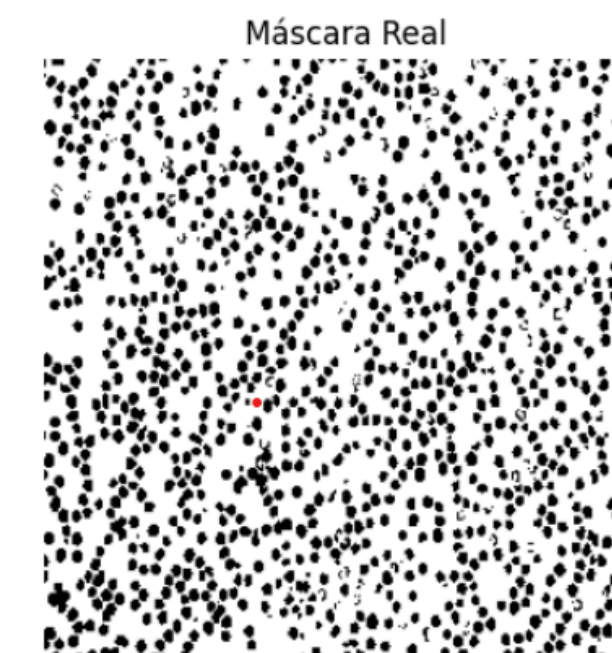
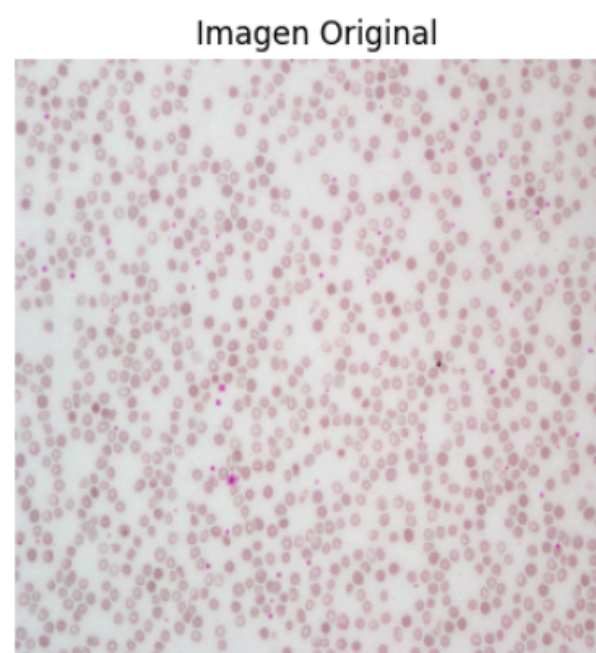
El modelo UNet preentrenado alcanzó un coeficiente Dice promedio de 0.9618, lo que indica una alta superposición entre la máscara predicha y la máscara real.

La distancia de Hausdorff promedio fue de 5.14 px, reflejando una precisión geométrica adecuada en los bordes de los eritrocitos segmentados

Evaluación por imagen — Coeficiente Dice



Evaluación — Distancia de Hausdorff
Promedio: 5.14 px



Resultados

Vector de características de las imagenes

- Para el vector se creo un nuevo dataset con el complemento de la mascara predicha de la U-net y la imagen real
- La red neuronal convolucional generó un vector de características de 128 dimensiones,
- Para identificar las características más influyentes, se aplicó el Análisis de Componentes Principales (PCA), permitiendo determinar la contribución de cada característica a la variabilidad de los datos y facilitar su interpretación.

118	119	120	121	122	123	124	125	126	127
0.034	-0.069209	-0.101596	-0.065700	0.084510	0.109541	0.096503	-0.108887	0.085264	-0.0341
0.050	0.130209	0.071595	0.073673	0.071369	0.048798	0.082397	0.054336	0.113509	0.1169
0.065	0.009659	0.003790	0.192780	0.112414	-0.057469	0.000922	-0.001007	-0.033819	0.1907
0.089	-0.069961	0.022165	0.033491	0.175002	0.021179	0.103676	0.042854	-0.030421	0.0881
0.064	-0.120329	-0.098277	0.173592	-0.081102	0.035260	-0.092239	0.001252	0.071615	-0.0426
...
0.047	0.026159	0.002298	-0.039651	0.019237	0.119391	0.030440	0.029463	0.004082	-0.0316
0.080	0.012686	0.009544	-0.004188	-0.020072	-0.370736	0.011820	0.005929	0.019267	0.0011
0.054	0.002972	0.003199	0.008991	-0.018294	0.699231	-0.012480	-0.016715	0.021036	-0.0229
0.041	0.018695	0.009243	0.004201	-0.000615	-0.295161	-0.016639	-0.006557	0.008629	-0.0545
0.029	-0.002903	0.001202	0.025374	0.010403	-0.088210	-0.003521	-0.003434	0.001725	-0.0145

Vector de las 128 características

Imagen filtrada

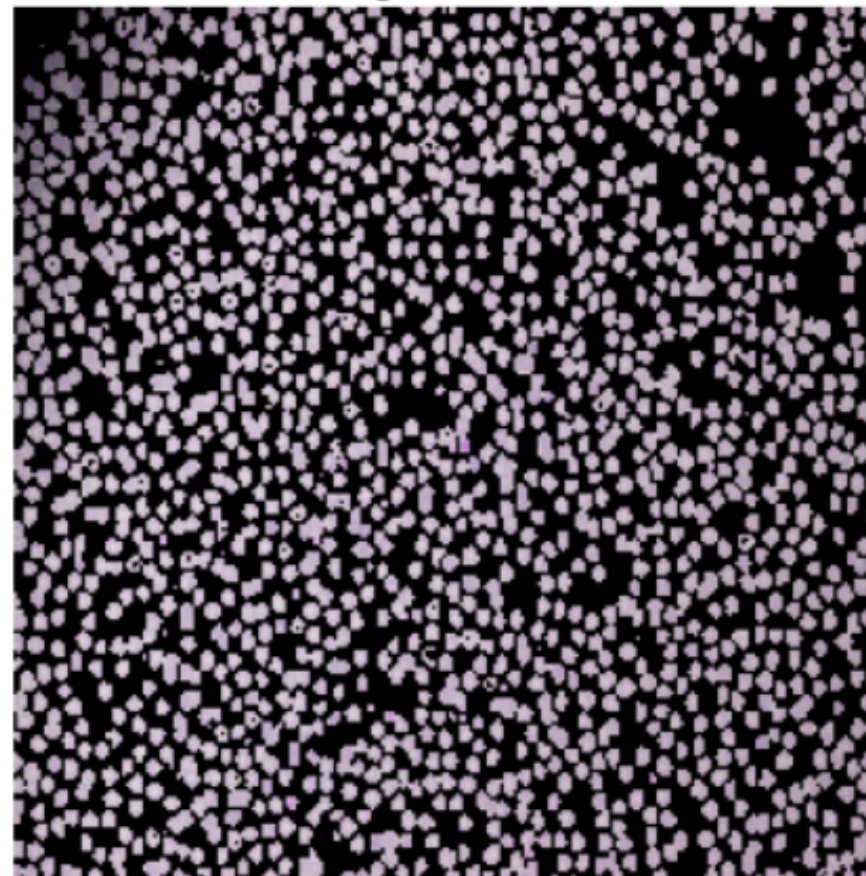
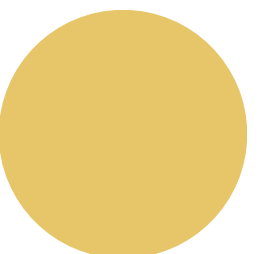


Imagen del nuevo dataset



Resultados

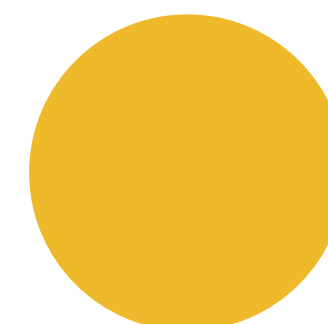
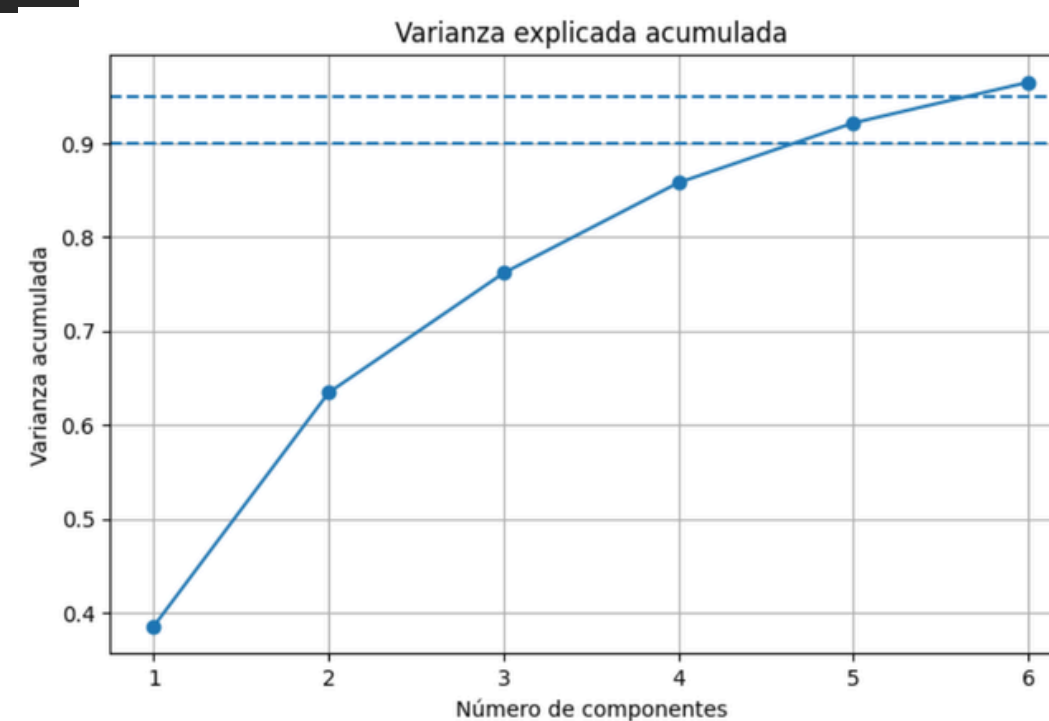
Vector de características de datos hematológicos

Vector de características obtenido con PCA para reducción de dimensionalidad.

Seleccionó automáticamente el número mínimo de componentes necesarios para conservar el 95% de la varianza total del conjunto de datos

Componente	Varianza explicada	Varianza acumulada
0	1	0.385484
1	2	0.634339
2	3	0.761695
3	4	0.857959
4	5	0.921635
5	6	0.964940

Varianza acumulada final: 96.49 %

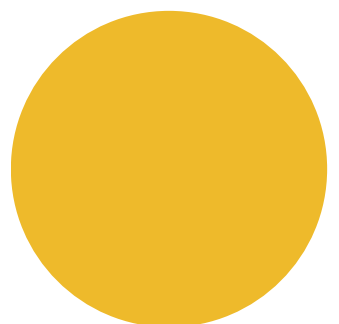


Resultados | Fusión Temprana

index	filepaths	labels	ids	WBC	RBC	HGB	HCT	MCV	MCH	MCHC	PLT	label	diagnosis
0	/content/drive/MyDrive/Cristi/Final_dataset/001_a.png	Anemica	001_a	8.49	4.37	10.1	29.9	68.4	23.1	33.8	376.0	1	1
1	/content/drive/MyDrive/Cristi/Final_dataset/001_h.png	Healthy	001_h	7.39	4.65	13.5	41.9	90.1	29.0	32.2	331.0	0	0
2	/content/drive/MyDrive/Cristi/Final_dataset/002_a.png	Anemica	002_a	9.16	4.94	12.3	35.6	72.1	24.9	34.6	232.0	1	1
3	/content/drive/MyDrive/Cristi/Final_dataset/002_h.png	Healthy	002_h	10.93	4.41	11.6	34.4	78.0	26.3	33.7	251.0	0	0
4	/content/drive/MyDrive/Cristi/Final_dataset/003_a.png	Anemica	003_a	21.29	4.46	10.8	32.9	73.8	24.2	32.8	501.0	1	1

Estrategia de concatenación de los vectores de características obtenidos de las salidas de ambos PCA's

Cada fila representa un paciente, cada columna representa una variable.





Gracias por su
atención