

Clasificación automática de células en muestras de citología oral mediante diferentes métodos de aprendizaje profundo

Automatic cell classification in oral cytology samples using different deep learning methods

Juana Iris Pérez Salas¹ <https://orcid.org/0009-0000-7734-8978>

Miriela Escobedo Nicot^{2*} <https://orcid.org/0000-0003-1103-3169>

Mildred Caridad Medrano Robert³ <https://orcid.org/0009-0003-8308-1028>

Wilkie Delgado Font⁴ <https://orcid.org/0000-0003-1431-2016>

Liuba González Espangler⁵ <https://orcid.org/0000-0002-2918-462X>

¹Universidad de Guantánamo; Departamento de Informática. Beneficencia entre 11 y 12 norte #71A.
juanairisperezsalas@gmail.com

²Universidad de Oriente; Dirección de Informatización. Avenida Las Américas entre Calle L y E.
miri@uo.edu.cu

³Hospital Dr. Ambrosio Grillo Portuondo. Carretera Central KM 21 y medio, Melgarejo, El Cobre.
mildredcaridadm@gmail.com

⁴Universidad de Oriente; Departamento de Ciencias de la Computación. Patricio Lumumba No. 507 entre Avenida Las Américas y Calle 1ra. wilkie.delgado@gmail.com

⁵ Facultad de Estomatología Santiago de Cuba, Sánchez Hechavarría esq Plácido,
liuba.gonzalez@infomed.sld.cu

*Autor para la correspondencia. (miri@uo.edu.cu)

RESUMEN

El cáncer oral presenta una alta mortalidad debido a diagnósticos tardíos, siendo la citología exfoliativa oral una herramienta clave no invasiva para la detección temprana. Sin embargo, su análisis manual es subjetivo y propenso a errores. Este estudio evalúa el uso del aprendizaje profundo para clasificar células de citología oral en tres categorías: anormales (asociadas a malignidad), sanguíneas (indicadoras de hemorragia) y saludables. Se examinaron seis modelos pre-entrenados: GhostNet-100, EfficientNet-Lite4, Inception-ResNet-v2, CoaTNet-0, MaxViT-Tiny y ViT Base Patch16, y su combinación mediante ensamble, utilizando el conjunto de datos UFSC OCPap con un total de 1934 imágenes. Para optimizar el rendimiento, se implementaron técnicas como el aumento de datos (rotaciones, volteos), Focal Loss y el optimizador SAM para manejar desbalances de clases. Los resultados demostraron que el ensamble superó a los modelos individuales, alcanzando un 84.03% de exactitud balanceada y un F1 del 88% para células anormales (sensibilidad del 87%). Este enfoque mejoró significativamente el estado del arte al superar en un 12% el F1 para células sanguíneas frente a estudios previos. La arquitectura Inception-ResNet-v2 destacó individualmente (83.31% exactitud), mientras que GhostNet-100 ofreció equilibrio entre eficiencia y precisión para entornos con recursos limitados. Estos resultados validan el potencial de los sistemas basados en aprendizaje profundo para optimizar el diagnóstico de cáncer oral, reduciendo falsos negativos en lesiones tempranas y proporcionando una herramienta objetiva de apoyo clínico.

Palabras clave: aprendizaje profundo; citología oral; cáncer oral; clasificación de imágenes.

ABSTRACT

Oral cancer has a high mortality rate due to late diagnosis, with oral exfoliative cytology being a key non-invasive tool for early detection. However, its manual analysis is subjective and prone to errors. This study evaluates the use of deep learning to classify oral cytology cells into three categories: abnormal (associated with malignancy), blood (indicative of hemorrhage), and healthy. Six pre-trained models were examined: GhostNet-100, EfficientNet-Lite4, Inception-ResNet-v2, CoaTNet-0, MaxViT-Tiny, and ViT Base Patch16,

and their combination through an ensemble, using the UFSC OCPap dataset with a total of 1934 images. To optimize performance, techniques such as data augmentation (rotations, flips), Focal Loss, and the SAM optimizer were implemented to handle class imbalances. The results showed that the ensemble outperformed individual models, achieving a balanced accuracy of 84.03% and an F1 of 88% for abnormal cells (sensitivity of 87%). This approach significantly improved the state-of-the-art by surpassing the F1 for blood cells by 12% compared to previous studies. The Inception-ResNet-v2 architecture stood out individually (83.31% accuracy), while GhostNet-100 offered a balance between efficiency and precision for resource-limited environments. These results validate the potential of deep learning-based systems to optimize the diagnosis of oral cancer, reducing false negatives in early lesions and providing an objective clinical support tool.

Keywords: deep learning; oral cytology; image classification; oral cancer; computer-aided diagnosis.

Recibido: 6/10/2025

Aceptado: 20/02/2026

Publicado: 01/04/2026

Introducción

El cáncer oral constituye un problema de salud pública a nivel mundial, caracterizado por una alta mortalidad, principalmente atribuida a su diagnóstico tardío. Cada año se reportan aproximadamente 380 000 nuevos casos y más de 180 000 muertes por esta enfermedad (Bray et al., 2024). Esta neoplasia puede afectar cualquier región de la cavidad oral y, cuando se diagnostica en estadios avanzados, requiere tratamientos agresivos que deterioran significativamente la calidad de vida de los pacientes (Warnakulasuriya, 2009). En consecuencia, la detección temprana resulta esencial para mejorar el pronóstico y reducir la mortalidad asociada.

La citología exfoliativa oral se ha consolidado como un procedimiento no invasivo para la detección de lesiones potencialmente malignas. Este método permite recolectar células epiteliales mediante técnicas como la biopsia por cepillado y analizarlas microscópicamente sin necesidad de intervención quirúrgica (Babshet et al., 2011). Sin embargo, la interpretación manual de estas muestras puede ser subjetiva y propensa a errores, lo que limita la precisión diagnóstica (Diniz Freitas et al., 2004). Ante este desafío, la inteligencia artificial (IA) se ha convertido en una alternativa prometedora para automatizar la clasificación citológica y aumentar la objetividad del proceso.

En particular, las redes neuronales profundas han mostrado un desempeño sobresaliente en la clasificación de imágenes médicas, gracias a su capacidad para extraer características complejas y reconocer patrones asociados a patologías (Litjens et al., 2017). Diversos estudios recientes han aplicado arquitecturas de aprendizaje profundo a la citología oral, logrando avances importantes en el diagnóstico temprano. Por ejemplo, Matias et al. (2021) integraron Mask R-CNN para segmentación y una red convolucional para clasificación, alcanzando un F1 cercano a 0.90. Sukegawa et al. (2022) exploraron optimizadores avanzados, como Sharpness-Aware Minimization (SAM), elevando la AUC hasta 0.93 y mejorando la estabilidad del entrenamiento. Otros trabajos han buscado modelos ligeros para entornos clínicos con recursos limitados (Kupas & Harangi, 2022), o han adaptado Vision Transformers para capturar relaciones espaciales de largo alcance, obteniendo métricas competitivas (Hörst et al., 2024). Finalmente, propuestas multimodales como CAFNet han demostrado que la integración de distintas fuentes de información óptica puede superar incluso el rendimiento humano en la detección de células malignas (Lian et al., 2025).

A pesar de estos avances, persisten limitaciones importantes en la literatura, entre ellas la escasa evaluación de arquitecturas modernas como transformers y modelos híbridos, el manejo insuficiente del desbalance de clases y la poca exploración de técnicas para mejorar la robustez de la clasificación, como el uso de ensambles de modelos. Estas brechas evidencian la necesidad de enfoques más integrales que aprovechen las capacidades de arquitecturas recientes y estrategias avanzadas de entrenamiento.

En este contexto se desarrolla el siguiente trabajo, que tiene como objetivo evaluar el desempeño de arquitecturas modernas pre-entrenadas: GhostNet-100 (Han et al., 2020), EfficientNet-Lite4 (Tan & Le, 2019), Inception-ResNet-v2 (Szegedy et al., 2016), (CoaTNet-0 Dai et al., 2021), MaxViT-Tiny (Tu et al.,

2022) y ViT Base Patch16 (Dosovitskiy et al., 2021), para la clasificación automática de células en imágenes de citología exfoliativa oral. Se seleccionaron estos seis modelos porque integran diferentes estrategias de diseño orientadas a optimizar la precisión, por la eficiencia computacional y la capacidad de generalización. Estas arquitecturas abarcan desde redes convolucionales ligeras para entornos con recursos limitados, hasta enfoques híbridos y transformadores que capturan patrones de mayor complejidad. Además de evaluar individualmente el rendimiento de cada modelo, se desarrolló un ensamble de los seis modelos seleccionados, con el objetivo de combinar sus fortalezas y mejorar la robustez y precisión del sistema de clasificación final.

Métodos o Metodología Computacional

Conjunto de datos

El conjunto de datos empleado en este estudio corresponde al UFSC OCPap: Papanicolaou Stained Oral Cytology Dataset v1, descrito originalmente por Matias et al. (2021). El mismo consta de imágenes obtenidas de dos portaobjetos de cepillado bucal diagnosticado con cáncer y teñido mediante la técnica de Papanicolaou, proporcionados por el Centro Odontológico del Hospital Universitario Professor Polydoro Ernani de São Thiago de la Universidade Federal de Santa Catarina, Brasil (HU-UFSC).

Las imágenes de 256×256 píxeles centrados en cada núcleo identificado, fueron catalogadas en cada una de las siguientes categorías: “núcleo epitelial anormal”, “núcleo epitelial saludable”, “núcleo fuera de foco”, “núcleo de célula sanguínea” y “núcleo en división”, además del fondo. Las mismas con sus respectivas máscaras, se dividieron en subconjuntos de entrenamiento, validación y prueba, siguiendo la proporción 70 %: 15 %: 15 %. (Tabla 1). No obstante, en el presente trabajo no se consideraron todas las categorías originales del conjunto de datos. Conforme a las observaciones de Matias et al. (2021), se excluyeron tanto las imágenes clasificadas como “fuera de foco” como aquellas correspondientes a “núcleos en división”. Las primeras pueden corresponder a cualquier otra clase, lo que introduce ambigüedad en el proceso de clasificación; las segundas presentan una representación numérica insuficiente para entrenar eficazmente los

modelos. De acuerdo con los autores, incluir estas clases podría inducir un rendimiento deficiente y sesgos en la clasificación. Como resultado de esta depuración, el presente estudio trabaja únicamente con las clases “anormal”, “saludable” y “sangre” (Fig. 1).

Tabla 1 - Distribución del conjunto original de imágenes.

Clases	Entrenamiento	Validación	Prueba
Anormal	1324	233	251
Divisor	27	7	6
Fuera de foco	651	137	122
Sangre	202	49	38
Saludable	900	187	153

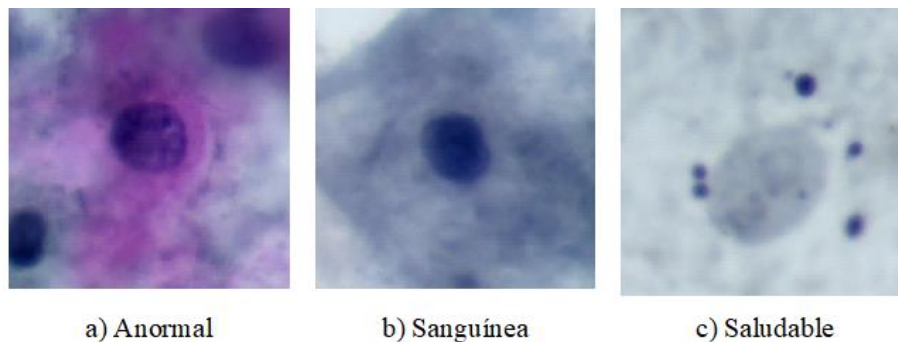


Fig. 1 - Clasificación de células de citología exfoliativa oral.

Para asegurar una comparación directa y rigurosa con los resultados reportados por (Matias et al., 2021), se empleó exactamente el mismo subconjunto depurado y la misma partición de datos definida en su estudio, garantizando así la validez de la evaluación comparativa del desempeño de los modelos.

En el conjunto de entrenamiento, se aplicaron varias técnicas de aumento de datos que incluyen: recorte aleatorio con redimensionamiento a 224×224 píxeles, rotación aleatoria de hasta 15 grados, traslaciones aleatorias, volteo horizontal aleatorio, y desenfoque gaussiano. Estas transformaciones buscan simular variaciones que pueden presentarse en escenarios reales y así mejorar la capacidad del modelo para

generalizar sobre datos no vistos (Shorten & Khoshgoftaar, 2019). Después de estas transformaciones, las imágenes son convertidas a tensores, y normalizadas utilizando los valores de media y desviación estándar del conjunto ImageNet: con los valores de media [0.485, 0.456, 0.406] y de desviación estándar [0.229, 0.224, 0.225], como se recomienda en modelos pre-entrenados.

Para los conjuntos de validación y prueba, se aplicó una transformación más conservadora consistente en el redimensionamiento directo a 224×224 píxeles seguido de la conversión a tensores y normalización. Esto permite evaluar el desempeño del modelo sobre datos sin alteraciones significativas que podrían influir en la predicción.

Arquitecturas

Las características principales de cada una de estas arquitecturas se describen a continuación (Tabla 2).

Tabla 2 - Redes pre-entrenadas seleccionadas para la clasificación de imágenes de citología exfoliativa oral.

Modelo	Año	Parámetros (M)	Tipo de arquitectura	Precisión Top-1 (%)
GhostNet-100	2020	5.2	CNN ligera	75.7
EfficientNet-Lite4	2020	13.0	CNN optimizada	80.4
Inception-ResNet-v2	2016	55.9	CNN híbrida	80.3
CoaTNet-0	2021	27.4	CNN + Transformer	81.6
MaxViT-Tiny	2022	30.9	Atención + CNN	83.6
ViT Base Patch16	2020	86.0	Transformer puro	84.0

Entrenamiento

El entrenamiento de las arquitecturas pre-entrenadas se realizó utilizando Python 3.11 y PyTorch, con el apoyo de la biblioteca timm para cargar los pesos pre-entrenados en ImageNet y adaptar cada modelo a un problema de clasificación con tres clases: 0 – anormal, 1 – sanguínea, y 2 – saludable.

En todos los casos, la capa de salida original fue reemplazada por un bloque de clasificación compuesto por una activación ReLU, una capa lineal y una función Softmax (Tabla 3). Este bloque permite transformar las características finales del modelo en probabilidades asociadas a cada clase.

Tabla 3 - Bloque de clasificación por arquitecturas.

Arquitectura	Capas de clasificación
GhostNet-100	ReLU → Linear (in=1280, out=3) → Softmax
EfficientNet-Lite4	ReLU → Linear (in=1280, out=3) → Softmax
Inception-ResNet-v2	ReLU → Linear (in=1536, out=3) → Softmax
CoaTNet-0	ReLU → Linear (in=768, out=3) → Softmax
MaxViT-Tiny	ReLU → Linear (in=512, out=3) → Softmax
ViT-Base-Patch16-224	ReLU → Linear (in=768, out=3) → Softmax

Para ajustar los pesos durante el entrenamiento, se utilizó la función de pérdida *Focal Loss*, definida como:

$$Fl(p_t) = -\alpha(1 - p_t)^\gamma \log(p_t) \quad (1)$$

Donde:

- p_t representa la probabilidad predicha para la clase verdadera.
- α corresponde a un factor de ponderación que controla la importancia relativa de cada clase.
- γ es un parámetro de ajuste que modula el efecto de enfoque.

Se emplearon los valores $\alpha = 1$ (sin ponderación entre clases) y $\gamma = 1$. Esta formulación penaliza con mayor fuerza los ejemplos difíciles (aquellos con p_t bajo), favoreciendo un aprendizaje más equilibrado en presencia de clases minoritarias o muestras ambiguas.

La optimización se realizó utilizando SAM, un optimizador de doble paso que favorece regiones planas en el paisaje de la superficie de la función de pérdida, lo que generalmente mejora la capacidad de generalización (Foret et al., 2021). Este método se aplicó sobre los optimizadores base SGD (del inglés *Stochastic Gradient Descent*) o AdamW, de acuerdo con la configuración óptima definida en la búsqueda de hiperparámetros. En particular, la combinación de SAM con SGD permite aprovechar la simplicidad y

estabilidad del descenso del gradiente estocástico, mientras se incorporan propiedad de SAM para evitar mínimos agudos. De forma complementaria, SAM con AdamW ofrece una actualización adaptativa de los parámetros junto con la penalización adecuada de los pesos, lo que mejora la estabilidad del entrenamiento.

Para favorecer una convergencia estable, la tasa de aprendizaje se programó aplicando un *warm-up* lineal durante las primeras 3–10 épocas, seguido de una estrategia de *Cosine Annealing*. Esta combinación, implementada mediante *SequentialLR* en PyTorch, permite iniciar con una tasa baja, incrementarla progresivamente y luego reducirla de manera cíclica, lo que contribuye a una mayor estabilidad y eficacia en el entrenamiento (Loshchilov & Hutter, 2016).

El sobreajuste se controló mediante un mecanismo de *Early Stopping* personalizado, que interrumpe el entrenamiento cuando no se observa mejora en la métrica de validación durante un número predefinido de épocas consecutivas o cuando la diferencia entre la precisión en entrenamiento y validación supera un umbral. En todos los casos, se conserva el mejor modelo alcanzado como punto de control.

Para favorecer la reproducibilidad del presente estudio, los experimentos se llevaron a cabo en una estación de trabajo equipada con un procesador Intel Xeon E5-2690 v4, 64 GB de memoria RAM y una GPU NVIDIA RTX 3090 con 24 GB de memoria. El entrenamiento de cada arquitectura requirió entre 35 y 70 minutos en promedio, dependiendo de la complejidad del modelo, mientras que la fase de inferencia en el conjunto de prueba demandó menos de 2 segundos por imagen en GPU y aproximadamente 65–80 ms por imagen en CPU. Dichos valores permiten estimar la factibilidad de implementación en entornos asistenciales con recursos computacionales heterogéneos

Valores de los hiperparámetros

La configuración óptima de hiperparámetros para cada arquitectura se determinó con la herramienta Optuna, empleando un *MedianPruner* para interrumpir tempranamente los experimentos con bajo desempeño (Akiba et al., 2019). Se exploraron los siguientes rangos:

1) Optimizador y regularización:

- a) Tasa de aprendizaje: $lr \in [10^{-5}, 10^{-3}]$
- b) Peso de regularización: $wd \in [10^{-6}, 10^{-4}]$
- c) Rango SAM: $\rho \in \{0.05, 0.1\}$
- d) Optimizador base: $optim \in \{SGD, AdamW\}$.

2) Esquema de tasa de aprendizaje:

- a) Épocas de warm-up: $warmup_epochs \in [3, 10]$
- b) Factor inicial: $start_factor \in [0.01, 0.2]$
- c) Épocas totales: $total_epochs \in [warmup+5, warmup+30]$.

Cada ensayo se entrenó de forma preliminar con un *Early Stopping* más estricto, con el fin de descartar rápidamente configuraciones ineficaces. Los hiperparámetros óptimos identificados para cada modelo se presentan a continuación (Tabla 4).

Tabla 4 - Hiperparámetros seleccionados tras la búsqueda con Optuna.

Arquitectura	Tasa de aprendizaje (lr)	Peso de regularización (wd)	Rango SAM (ρ)	Optimizador	Épocas de warm-up	Factor inicial	Épocas totales
GhostNet-100	9.2e-4	9.2e-5	0.10	AdamW	9	0.19	19
EfficientNet-Lite4	5.1e-3	2.7e-4	0.10	SGD	4	0.05	33
Inception-ResNet-v2	3.0e-3	7.5e-4	0.05	SGD	6	0.15	25
CoaTNet-0	3.0e-4	5.0e-5	0.10	AdamW	7	0.08	14
MaxViT-Tiny	6.8e-4	1.5e-5	0.05	AdamW	10	0.07	10
ViT-Base-Patch16-224	9.4e-5	1.3e-5	0.10	AdamW	5	0.11	22

Una vez seleccionados los hiperparámetros óptimos, cada modelo se entrenó durante un máximo de 100 épocas o hasta la activación del *Early Stopping*. En cada época se registraron las métricas de pérdida y precisión tanto en entrenamiento como en validación, y se conservó el mejor modelo para su posterior

evaluación en el conjunto de prueba. El número de parámetros entrenables para cada arquitectura se presenta en la Tabla 5.

Tabla 5 - Cantidad de parámetros entrenables.

Arquitectura	Número de parámetros entrenables (Millones)
GhostNet-100	3.9
EfficientNet-Lite4	11.7
Inception-ResNet-v2	54.3
CoaTNet-0	26.6
MaxViT-Tiny	28.5
ViT-Base-Patch16-224	85.8

Marco experimental

Se utilizó la matriz de confusión para evaluar los resultados, con las siguientes medidas: sensibilidad (TPR), precisión (P) y puntuación F1 (F1) para cada clase (Escobedo et al., 2024). A continuación, se describen estas medidas.

$$TPR = \frac{TP}{TP + FN} \quad (2)$$

$$P = \frac{TP}{TP + FP} \quad (3)$$

$$F1 = \frac{2(P \cdot TPR)}{P + TPR} \quad (4)$$

Donde:

- TP se corresponde a los verdaderos positivos.

- FP se corresponde a los falsos positivos.
- FN se corresponde a los falsos negativos.

Además, se empleó la exactitud balanceada (Acc_B) el cual define el promedio de las sensibilidades de cada clase, compensando el efecto de clases mayoritarias:

$$Acc_B = \frac{1}{n} \sum_{i=1}^n TPR_i \quad (5)$$

Resultados y discusión

El desempeño de los modelos se evaluó utilizando un conjunto de prueba independiente, compuesto por 442 imágenes inéditas (251 anormales, 38 sanguíneas y 153 saludables, Tabla 6). Esta configuración asegura una evaluación representativa y realista del rendimiento de los modelos en condiciones clínicas simuladas.

Tabla 6 - Composición del conjunto de prueba.

Clase	Etiqueta	Prueba
Anormal (A)	0	251
Sangre (SN)	1	38
Saludable (SL)	2	153

En este estudio se emplearon métricas como la exactitud balanceada (Acc_B) y el F1 por clase, cuya interpretación trasciende lo meramente técnico. En particular, el F1 de la clase anormal resulta crítico desde el punto de vista clínico, ya que un valor elevado implica reducir falsos negativos en lesiones potencialmente malignas, lo cual es esencial para la detección temprana de cáncer oral. De igual modo, la Acc_B ofrece una medida global que compensa el desbalance de clases, garantizando una evaluación más justa del rendimiento del modelo en todas las categorías.

El análisis del proceso de entrenamiento revela patrones consistentes en todas las arquitecturas. Se observa una disminución sostenida en las curvas de pérdida tanto para entrenamiento como validación a lo largo de las épocas, mientras que las curvas de exactitud presentan una tendencia ascendente convergente (Fig. 2 y Fig. 3). Hay que resaltar que todas las arquitecturas alcanzaron estabilidad antes de la época 40. con tiempos de convergencia variables según su complejidad, validando así los criterios implementados.

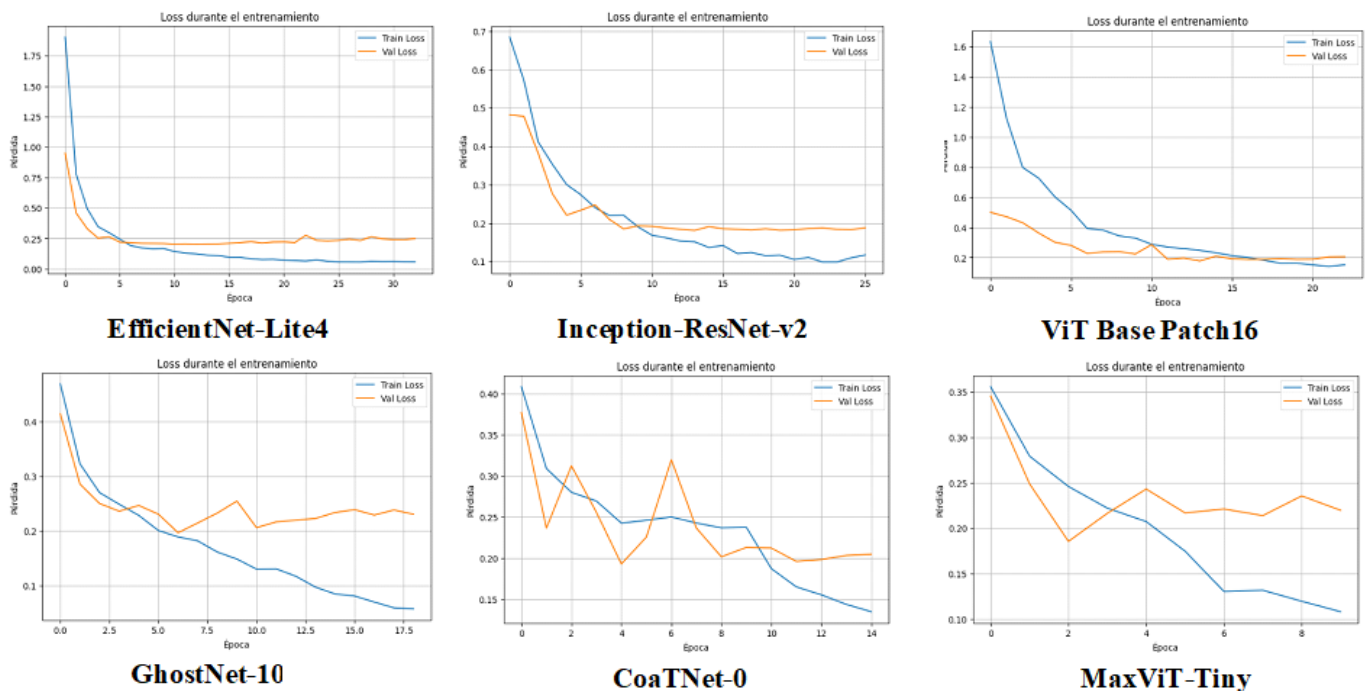


Fig. 2 - Curvas de pérdida durante el entrenamiento de los modelos evaluados.

Se observa que, en la mayoría de modelos, la diferencia entre pérdidas de entrenamiento y validación se mantuvo por debajo del 10% en fases avanzadas, confirmando que la optimización con SAM logró encontrar regiones planas en el espacio de pérdidas que favorecen la generalización. Simultáneamente, la brecha en exactitud se contuvo sistemáticamente dentro del umbral del 15% garantizado por el *Early Stopping*, con valores típicos inferiores al 10%. Este comportamiento controlado demuestra la efectividad sinérgica de las técnicas implementadas: SAM redujo la sensibilidad a mínimos locales, el *focal loss* contrarrestó el desbalance de clases, el aumento de datos enriqueció la variabilidad muestral, y la

búsqueda de hiperparámetros con Optuna optimizó la adaptabilidad de cada arquitectura a las particularidades del conjunto.

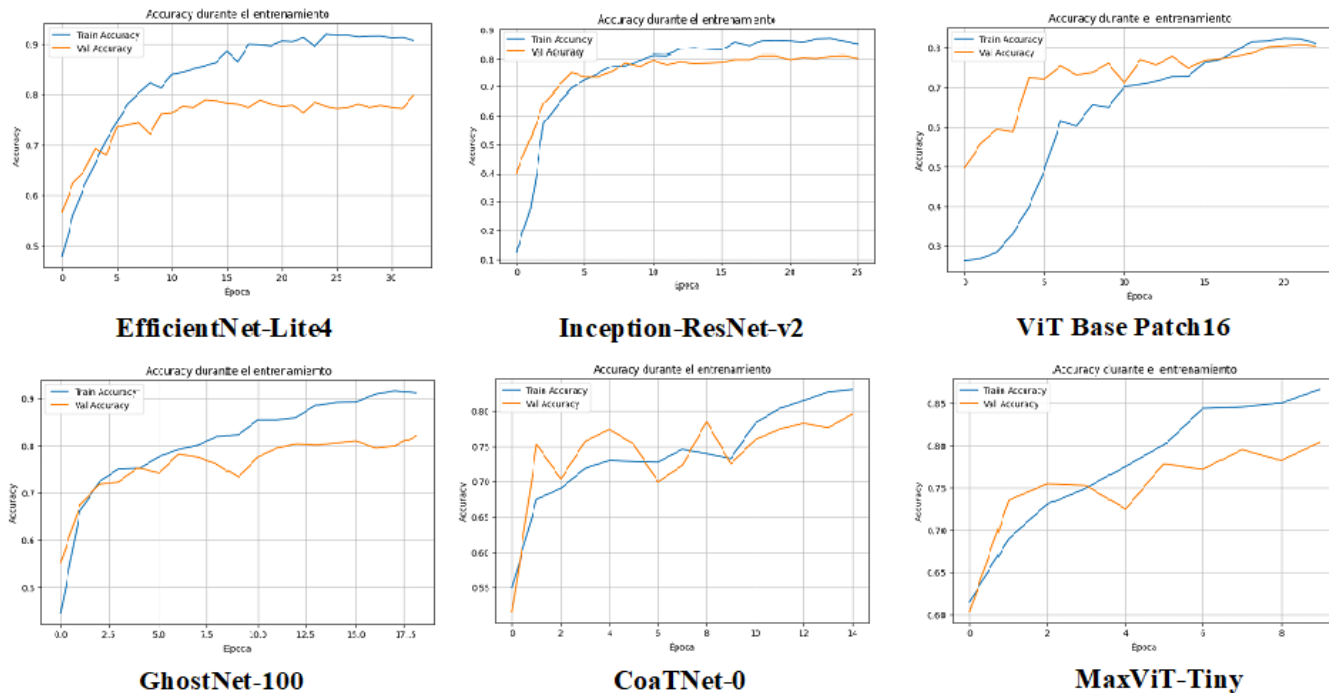


Fig. 3 - Curvas de evolución de la exactitud durante el entrenamiento.

Para el ensamble, al tratarse de una combinación de modelos previamente entrenados, no se generaron curvas propias. Sin embargo, la estabilidad observada en los componentes individuales fundamentó su rendimiento robusto en la fase de prueba. Por otra parte, en la tabla 7 se muestran los resultados de las matrices de confusión en la etapa de prueba para todos los métodos valorados.

El análisis de los resultados revela patrones significativos vinculados a la complejidad morfológica y el desbalance de clases. Como se observa en las matrices de confusión, la clase SN presenta la mayor confusión relativa a su tamaño muestral, particularmente en su identificación errónea como SL. Por ejemplo, el modelo ViT Base Patch16 clasificó incorrectamente 4 muestras sanguíneas como saludables, mientras que MaxViT-Tiny cometió 7 errores en esta categoría. Este fenómeno se atribuye a dos factores

principales: la escasa representación de la clase sanguínea (solo 38 muestras, equivalente al 8.6% del conjunto) y la similitud morfológica entre células hemolizadas y tejido sano en preparaciones citológicas.

Tabla 7 - Matrices de confusión por modelos para la prueba.

	EfficientNet-Lite4			Inception-ResNet-v2			ViT Base Patch16					
	A	SN	SL	A	SN	SL	A	SN	SL			
A	214	6	31	206	10	35	222	9	20			
SN	8	27	3	3	32	3	5	29	4			
SL	24	3	126	22	3	128	32	3	118			
	GhostNet-100			CoaTNet-0			MaxViT-Tiny			Ensamble		
	A	SN	SL	A	SN	SL	A	SN	SL	A	SN	SL
A	215	3	33	218	7	26	214	5	32	218	6	27
SN	6	29	3	3	32	3	8	23	7	5	31	2
SL	25	2	126	29	6	118	23	3	127	22	3	128

Los resultados cuantitativos expuestos (Tabla 8) demuestran que el ensamble alcanza el mejor desempeño global con un 84.03% de exactitud balanceada, superando a todos los modelos individuales. Esta superioridad se refleja particularmente en la clase A, donde logra un F1 del 88% (precisión: 89%, sensibilidad: 87%), métrica crucial para aplicaciones diagnósticas al minimizar falsos negativos en anomalías celulares. Para la clase SN, aunque persisten desafíos, el ensamble consigue una mejora del 12-13% en F1 respecto a estudios previos (79% vs. 66-67% en ResNet), resultado de la complementariedad entre arquitecturas como Inception-ResNet-v2 (84% sensibilidad) y CoaTNet-0 (84% sensibilidad).

Como se ilustra, la comparación de la exactitud balanceada confirma la superioridad del ensamble (84.03%) frente a arquitecturas individuales (Fig.4). Destacan Inception-ResNet-v2 (83.31%) y CoaTNet-0 (82.73%) como los modelos individuales más robustos, mientras que MaxViT-Tiny (76.26%) muestra limitaciones en el manejo del desbalance de clases. Esta jerarquía de rendimiento se correlaciona con la capacidad de los modelos para extraer características discriminativas: las arquitecturas híbridas (Inception-ResNet-v2) y transformadores (CoaTNet-0) superan a redes puramente convolucionales en la identificación de patrones sutiles en células anormales.

Tabla 8 - Valores porcentuales de precisión, sensibilidad, F1 y exactitud balanceada del proceso resultante de la prueba.

EfficientNet-Lite4	A	SN	SL	Inception-ResNet-v2	A	SN	SL	ViT Base Patch16	A	SN	SL
P	87	75	79	P	89	71	77	P	86	71	83
TPR	85	71	82	TPR	82	84	84	TPR	88	76	77
F1	86	73	81	F1	85	77	80	F1	87	73	80
Acc_B	79.55			Acc_B	83.31			Acc_B	80.63		
GhostNet-100	A	SN	SL	CoaTNet-0	A	SN	SL	MaxViT-Tiny	A	SN	SL
P	87	85	78	P	87	71	80	P	87	74	77
TPR	86	76	82	TPR	87	84	77	TPR	85	61	83
F1	87	81	80	F1	87	77	79	F1	86	67	80
Acc_B	81.44			Acc_B	82.73			Acc_B	76.26		
Ensamble					A	SN	SL				
					P	89	78	82			
					TPR	87	82	84			
					F1	88	79	83			
					Acc_B	84.03					

El análisis de los errores muestra que la mayoría de las equivocaciones se concentran en la distinción entre las clases sanguínea y saludable, lo cual coincide con la naturaleza morfológica de estas células en las preparaciones citológicas. En particular, las células sanguíneas con hemólisis parcial pueden presentar contornos celulares difusos y una coloración tenue que se asemeja a células epiteliales con tinción débil, lo que conduce a una mayor tasa de falsos negativos en la clase sanguínea. También se observaron errores en la clasificación de células anormales con núcleos de tamaño reducido o con pleomorfismo leve, que fueron confundidas con células saludables, lo cual es clínicamente relevante porque estos casos corresponden precisamente a lesiones en estadios tempranos, donde la detección es más difícil y más importante. Por otro lado, se identificaron algunos falsos positivos en la clase anormal asociados a artefactos de tinción y a núcleos superpuestos, donde los modelos basados en atención interpretaron el aumento de densidad nuclear como un indicador de malignidad. Estos resultados sugieren que el sistema tiende a sobreestimar patrones nucleares intensos y subestimar variaciones sutiles en citoplasma y bordes celulares, lo que indica la necesidad de incorporar mecanismos de interpretabilidad.

Desde la perspectiva clínica, los errores de clasificación tienen implicaciones relevantes para el diagnóstico y el manejo del paciente. Los falsos negativos en la clase anormal representan el riesgo más crítico, pues pueden conducir a la no identificación temprana de lesiones potencialmente malignas, retrasando el inicio del tratamiento y disminuyendo la probabilidad de control de la enfermedad. Por otro lado, los falsos positivos, especialmente en la clasificación de células saludables como anormales, pueden generar ansiedad en el paciente, indicar estudios adicionales innecesarios e incrementar la carga asistencial. En el caso de las células sanguíneas, los falsos negativos pueden ocultar signos de inflamación o trauma local, afectando la interpretación integral del contexto citopatológico. Por tanto, maximizar la sensibilidad en la detección de células anormales y equilibrarla con una adecuada especificidad es esencial para que estos sistemas funcionen de manera segura como apoyo al diagnóstico clínico.

La comparación con la literatura, presentada en la Tabla 9, evidencia avances frente al trabajo de Matias et al. (2021) con ResNet-34 y ResNet-50. El ensamble propuesto alcanzó mejoras del +2 % en células anormales (88 % vs. 86 %), +12 % en células sanguíneas (79 % vs. 67 %) y +7 % en células saludables (83 % vs. 76 %). Estas mejoras pueden atribuirse a la combinación de varios factores: la integración de modelos con sesgos complementarios, tanto convolucionales como basados en atención; la optimización mediante SAM, que contribuye a reducir el sobreajuste en clases minoritarias y la aplicación de técnicas avanzadas de regularización, como *Early Stopping* con monitoreo por clase.

Desde el punto de vista clínico, la interpretabilidad constituye un factor crítico para la adopción de sistemas de apoyo diagnóstico. Si bien el presente estudio no incorporó técnicas de explicabilidad, resultados previos en aplicaciones citológicas sugieren que mecanismos como Grad-CAM o capturas de atención en transformers permiten visualizar regiones nucleares y citoplasmáticas relevantes durante la toma de decisiones. La integración de estos mecanismos en estudios futuros habilitaría la validación por parte del citopatólogo, favoreciendo la trazabilidad diagnóstica y reduciendo la incertidumbre en casos ambiguos.

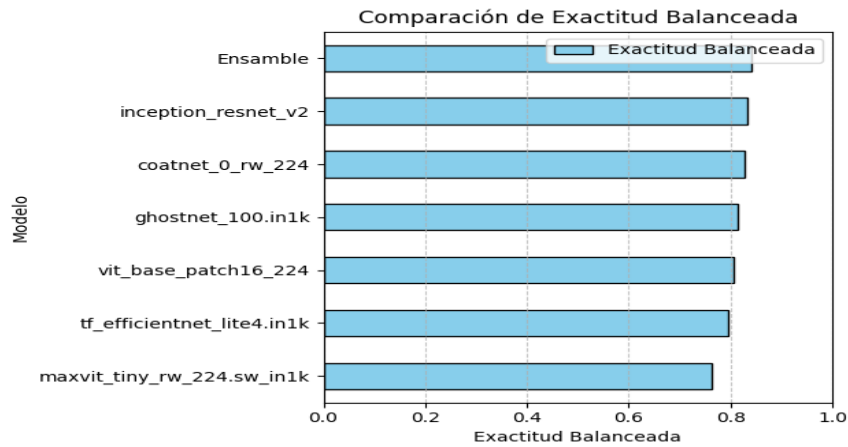


Fig. 4 - Comparación de los modelos según la exactitud balanceada.

Tabla 9 - Comparativa con métodos de la literatura en relación con la medida F1 por clase.

Método	F1 (%)		
	A	SN	SL
GhostNet-100	87	81	80
EfficientNet-Lite4	86	73	81
Inception-ResNet-v2	85	77	80
CoaTNet-0	87	77	79
MaxViT-Tiny	86	67	80
ViT Base Patch16	87	73	80
Ensamble	88	79	83
ResNet 34 Matias et al. (2021)	86	66	76
ResNet 50 Matias et al. (2021)	85	67	75

Viabilidad práctica en el contexto cubano

La implementación de sistemas de apoyo diagnóstico basados en aprendizaje profundo en el contexto cubano requiere considerar las condiciones reales de infraestructura tecnológica y disponibilidad de recursos

en los servicios de salud. El enfoque propuesto resulta especialmente adecuado para este escenario, ya que incluye modelos como GhostNet-100, cuya baja demanda computacional permite su ejecución en equipos estándar presentes en laboratorios de citología y centros de diagnóstico provinciales, sin necesidad de GPU de alto rendimiento. Para su despliegue, se requiere una infraestructura mínima compuesta por estaciones de trabajo con acceso a microscopía digital o cámaras acopladas a microscopios convencionales, así como conectividad local para el intercambio de resultados. La integración con los flujos clínicos actuales puede realizarse de manera gradual, incorporando el sistema como herramienta de apoyo al citopatólogo, donde este valide y corrija las predicciones generadas, evitando depender de una decisión automática. Además, la capacitación técnica del personal y la incorporación del software en las redes hospitalarias del Sistema Nacional de Salud (preferiblemente mediante soluciones offline o en servidores locales) contribuirían a garantizar sostenibilidad, seguridad de los datos y continuidad diagnóstica, favoreciendo su adopción sin generar cargas adicionales sobre la infraestructura existente.

Conclusiones

El presente estudio evaluó de manera sistemática el desempeño de diversas arquitecturas de aprendizaje profundo para la clasificación automática de células en imágenes de citología exfoliativa oral, demostrando que la combinación de modelos convolucionales, híbridos y basados en atención, junto con técnicas avanzadas de optimización y regularización, permite obtener resultados competitivos en un escenario caracterizado por el desbalance de clases y la complejidad morfológica de las muestras citológicas.

Los resultados obtenidos evidencian que el enfoque basado en ensamble alcanza un mejor equilibrio entre sensibilidad y precisión que los modelos individuales, particularmente en la identificación de células anormales, donde se logra un F1-score del 88 % y una sensibilidad del 87 %. Estas métricas resultan clínicamente relevantes, ya que contribuyen a reducir el riesgo de falsos negativos en lesiones potencialmente malignas, uno de los principales desafíos en el diagnóstico temprano del cáncer oral. Asimismo, se observaron mejoras sustanciales en la clasificación de células sanguíneas en comparación con

estudios previos, lo que sugiere que la integración de arquitecturas con sesgos complementarios favorece una mayor robustez del sistema.

Desde el punto de vista práctico, los resultados indican que la implementación de este tipo de investigación como herramientas de apoyo diagnóstico es técnicamente factible. Esto abre la posibilidad de una adopción progresiva en laboratorios de citología y centros de diagnóstico provinciales, siempre bajo un esquema de validación humana por parte del citopatólogo y sin sustituir el juicio clínico especializado. Como líneas futuras de investigación, se propone la incorporación de técnicas de explicabilidad que faciliten la interpretación de las predicciones por parte del personal médico; la evaluación del desempeño en condiciones operativas reales, incluyendo tiempos de respuesta y requisitos de integración con los procesos clínicos existentes y el análisis del impacto del uso de estas herramientas en la calidad y oportunidad del diagnóstico citológico.

En conjunto, este trabajo aporta evidencia técnica que respalda el potencial del aprendizaje profundo como apoyo al diagnóstico citológico oral y sienta bases metodológicas para el desarrollo de soluciones de inteligencia artificial adaptadas a las condiciones reales del sistema de salud cubano, priorizando la seguridad clínica, la sostenibilidad tecnológica y la integración progresiva en la práctica asistencial.

Referencias

Akiba, T., Sano, S., Yanase, T., Ohta, T., Y Koyama, M. *Optuna: A Next-Generation Hyperparameter Optimization Framework*. En: *Proceedings Of The 25th Acm Sigkdd International Conference On Knowledge Discovery & Data Mining*. Acm, 2019, P. 2623–2631.

Babshet, M., Et Al. *Efficacy Of Oral Brush Cytology In The Evaluation Of The Oral Premalignant And Malignant Lesions*. *Journal Of Cytology*, 2011, Vol. 28, No. 4, P. 165–172.

Bray, F., Et Al. *Global Cancer Statistics 2022: Globocan Estimates Of Incidence And Mortality Worldwide For 36 Cancers In 185 Countries*. *Ca: A Cancer Journal For Clinicians*, 2024, Vol. 74, No. 3, P. 229–263.

- Dai, Z., Et Al. *Coatnet: Marrying Convolution And Attention For All Data Sizes. Advances In Neural Information Processing Systems*, 2021, Vol. 34, P. 3965–3977.
- Diniz Freitas, M., Et Al. *Aplicaciones De La Citología Exfoliativa En El Diagnóstico Del Cáncer Oral. Medicina Oral, Patología Oral Y Cirugía Bucal (Ed. Impresa)*, 2004, Vol. 9, No. 4, P. 355–361.
- Dosovitskiy, A., Et Al. *An Image Is Worth 16x16 Words: Transformers For Image Recognition At Scale. Arxiv Preprint Arxiv:2010.11929*, 2020. [En Línea]. Disponible En: <https://doi.org/10.48550/Arxiv.2010.11929> [Consultado: 30 Septiembre 2025].
- Escobedo, M., Et Al. *Deep Learning Models For Automatic Morphological Evaluation Of Endothelial Cells. En Brazilian Congress On Biomedical Engineering. Cham: Springer Nature Switzerland*, 2024. P. 1106-1117.
- Foret, P., Kleiner, A., Mobahi, H., Y Neyshabur, B. *Sharpness-Aware Minimization For Efficiently Improving Generalization. Arxiv Preprint Arxiv:2010.01412*, 2020. [En Línea]. Disponible En: <https://arxiv.org/abs/2010.01412> [Consultado: 30 Septiembre 2025].
- Han, K., Et Al. *Ghostnet: More Features From Cheap Operations. En: Proceedings Of The Ieee/Cvf Conference On Computer Vision And Pattern Recognition. Ieee*, 2020. P. 1580–1589.
- Hörst, F., Et Al. *Cellvit: Vision Transformers For Precise Cell Segmentation And Classification. Medical Image Analysis*, 2024, Vol. 94, P. 103143.
- Kupas, D., Y Harangi, B. *Classification Of Pap-Smear Cell Images Using Deep Convolutional Neural Network Accelerated By Hand-Crafted Features. En: 2022 44th Annual International Conference Of The Ieee Engineering In Medicine & Biology Society (Embc). Ieee*, 2022. P. 1452–1455.
- Lian, W., Et Al. *Let It Shine: Autofluorescence Of Papanicolaou-Stain Improves Ai-Based Cytological Oral Cancer Detection. Computers In Biology And Medicine*, 2025, Vol. 185, P. 109498.
- Litjens, G., Et Al. *A Survey On Deep Learning In Medical Image Analysis. Medical Image Analysis*, 2017, Vol. 42, P. 60–88.
- Loshchilov, I., Y Hutter, F. *Sgdr: Stochastic Gradient Descent With Warm Restarts. Arxiv Preprint Arxiv:1608.03983*, 2016. [En Línea]. Disponible En: <https://arxiv.org/abs/1608.03983> [Consultado: 30 Septiembre 2025].

- Matias, A. V., Et Al. *Segmentation, Detection, And Classification Of Cell Nuclei On Oral Cytology Samples Stained With Papanicolaou*. *Sn Computer Science*, 2021, Vol. 2, No. 4, P. 285.
- Shorten, C., Y Khoshgoftaar, T. M. *A Survey On Image Data Augmentation For Deep Learning*. *Journal Of Big Data*, 2019, Vol. 6, No. 1, P. 1–48.
- Sukegawa, S., Et Al. *Effective Deep Learning For Oral Exfoliative Cytology Classification*. *Scientific Reports*, 2022, Vol. 12, No. 1, P. 13281.
- Szegedy, C., Et Al. *Inception-V4, Inception-Resnet And The Impact Of Residual Connections On Learning*. En: *Proceedings Of The Aaai Conference On Artificial Intelligence*. Aaai, 2017.
- Tan, M., Y Le, Q. V. *Efficientnet: Rethinking Model Scaling For Convolutional Neural Networks*. En: *Proceedings Of The 36th International Conference On Machine Learning*. Pmlr, 2019. P. 6105–6114.
- Tu, Z., Et Al. *Maxvit: Multi-Axis Vision Transformer*. En: *European Conference On Computer Vision*. Cham: Springer Nature Switzerland, 2022. P. 459–479.
- Warnakulasuriya, S. *Global Epidemiology Of Oral And Oropharyngeal Cancer*. *Oral Oncology*, 2009, Vol. 45, No. 4–5, P. 309–316.

Conflicto de interés

Los autores autorizan la distribución y uso de su artículo.

Contribuciones de los autores

- Conceptualización: Juana Iris Pérez Salas, Miriela Escobedo Nicot, Mildred Caridad Medrano Robert
- Curación de datos: Juana Iris Pérez Salas, Miriela Escobedo Nicot
- Análisis formal: Juana Iris Pérez Salas, Miriela Escobedo Nicot
- Adquisición de fondos: Miriela Escobedo Nicot
- Investigación: Juana Iris Pérez Salas
- Metodología: Miriela Escobedo Nicot, Juana Iris Pérez Salas, Mildred Caridad Medrano Robert, Wilkie Delgado Font
- Administración del proyecto: Miriela Escobedo Nicot, Wilkie Delgado Font, Liuba González Espangler
- Recursos: Miriela Escobedo Nicot, Mildred Caridad Medrano Robert, Wilkie Delgado Font

Software: Juana Iris Pérez Salas

Supervisión: Miriela Escobedo Nicot, Wilkie Delgado Font

Validación: Juana Iris Pérez Salas, Miriela Escobedo Nicot

Visualización: Juana Iris Pérez Salas, Miriela Escobedo Nicot

Redacción – borrador original: Juana Iris Pérez Salas, Miriela Escobedo Nicot

Redacción – revisión y edición: Juana Iris Pérez Salas, Miriela Escobedo Nicot