
This is the **published version** of the bachelor thesis:

Martínez Zamora, Luis. *Mejora de modelos de diagnóstico médico en colonoscopias mediante Data-Centric AI*. Treball de Final de Grau (Universitat Autònoma de Barcelona), 2026 (Enginyeria Informàtica)

This version is available at <https://ddd.uab.cat/record/334377>

under the terms of the  license.

Mejora de modelos de diagnóstico médico en colonoscopias mediante Data-Centric AI

Luis Martínez Zamora

30 de junio de 2026

Resumen– Este trabajo desarrolla un flujo automatizado basado en Data-Centric AI para mejorar la clasificación multiclase de pólipos colorrectales a partir de vídeos de colonoscopia e imágenes anotadas por especialistas. El flujo amplía y refina el conjunto de entrenamiento mediante la extracción selectiva de fotogramas, la detección de pólipos y la reducción de redundancia temporal, sin modificar la arquitectura de los clasificadores. Se evaluó con EfficientNetB0 y ViT-Small, manteniendo fijos los conjuntos de validación y test. El rendimiento mejoró en validación para ambas arquitecturas, pero no de forma consistente en un test externo multicéntrico. Los resultados muestran la utilidad del enfoque dentro del dominio de entrenamiento y la necesidad de incorporar mayor diversidad de centros y dispositivos.

Palabras clave– Clasificación multiclase, Colonoscopia, Data-Centric AI, Datos desequilibrados, Visión por computador.

Abstract– This work develops an automated Data-Centric AI workflow to improve the multiclass classification of colorectal polyps from colonoscopy videos and images annotated by specialists. The workflow expands and refines the training set through selective frame extraction, polyp detection, and temporal redundancy reduction, without modifying the classifier architectures. It was evaluated using EfficientNetB0 and ViT-Small while keeping the validation and test sets fixed. Validation performance improved for both architectures, but the improvement was not consistently reproduced on an external multicenter test set. The results show the usefulness of the approach within the training domain and the need to incorporate greater diversity of centers and acquisition devices.

Keywords– Colonoscopy, Computer Vision, Data-Centric AI, Imbalanced Data, Multiclass Classification.

1 INTRODUCCIÓN

EL cáncer de colon, también conocido como cáncer colorrectal, es una enfermedad que afecta al intestino grueso, la parte final del aparato digestivo. Este cáncer es el tercero más común a nivel mundial según la Organización Mundial de la Salud (OMS)[1], situándose además como la segunda causa de mortalidad oncológica. Por estos motivos, la prevención y detección de esta enfermedad es fundamental para reducir su impacto.

En la práctica clínica, el cribado del cáncer colorrectal

suele iniciarse mediante pruebas no invasivas, como los análisis de heces o sangre[2]. Cuando sus resultados indican la posible presencia de lesiones, se remite al paciente para que se realice una colonoscopia para realizar una evaluación directa y continuar el estudio diagnóstico.

La colonoscopia es un procedimiento médico que permite examinar el interior del intestino grueso mediante un colonoscopio[3]. La visualización en tiempo real de la mucosa facilita la detección y, cuando está indicado, la extirpación de pólipos durante la misma intervención. Por ello, constituye una herramienta central para la prevención y el diagnóstico del cáncer colorrectal.

Los pólipos colorrectales son lesiones que crecen sobre la mucosa del colon o del recto y no constituyen un grupo homogéneo[4]. En este trabajo se consideran cuatro categorías histológicas: adenomas, adenomas serrados sésiles, pólipos hiperplásicos y adenocarcinomas. Los adenomas son lesio-

- E-mail de contacte: lmzamora@gmail.com
- Menció: Computació
- Treball tutoritzat per: Yael Tudela (Ciències de la Computació)
- Curs 2025/26

nes neoplásicas precursoras de cancer colorectal (CRC); los adenomas serrados sésiles son otro tipo de polipo que pese a que puede derivar también en CRC, este sigue otro proceso y es necesario distinguirlo; los pólipos hiperplásicos son, por lo general, lesiones no neoplásicas de bajo potencial maligno; y los adenocarcinomas corresponden a lesiones malignas establecidas. Estas categorías pueden diferir en morfología, textura y relevancia clínica, lo que dificulta su diferenciación a partir de la imagen endoscópica.

Actualmente, la clasificación histológica definitiva de los pólipos se obtiene tras su extirpación y a través de el análisis patológico (AP) del tejido extirpado[2]. Sin embargo, durante la colonoscopia el especialista realiza una primera valoración visual de la lesión a partir de características como su tamaño, forma o textura. Esta valoración puede resultar compleja cuando distintas lesiones presentan rasgos similares o son de pequeño tamaño. Por ello, los sistemas de diagnóstico asistido por ordenador (*Computer-Aided Diagnosis*, CADx) pueden actuar como apoyo para estimar en tiempo real la categoría histológica de los pólipos observados durante la exploración[5].

Una formulación habitual para el diagnóstico óptico agrupa las lesiones en categorías neoplásicas y no neoplásicas[6, 7]. Dentro de las clases consideradas en este trabajo, los adenomas, los adenomas serrados sésiles y los adenocarcinomas pertenecen al primer grupo, mientras que los pólipos hiperplásicos se consideran lesiones no neoplásicas. Aunque esta clasificación resulta útil para determinadas decisiones clínicas, no recoge las diferencias entre las cuatro categorías. Por ello, este trabajo aborda la clasificación multiclase de todas ellas, aportando un nivel de detalle más adecuado para el diagnóstico.

Computacionalmente, esta transición introduce un problema crítico: las cuatro categorías presentan diferencias visuales sutiles y tienen una prevalencia clínica muy dispar. Esto puede generar un desbalanceo natural entre clases, donde la escasez de imágenes de las categorías menos comunes puede provocar que los clasificadores pierdan eficacia al aprender sus patrones representativos.

El desbalance en la representación de clases es un problema habitual en los datos médicos, asociado a factores como la baja prevalencia de determinadas patologías, los sesgos en la recogida de datos y las restricciones éticas y de privacidad[8]. En este trabajo, el problema resulta especialmente relevante en la clasificación de pólipos, donde algunas categorías son poco frecuentes y presentan diferencias visuales sutiles. Por tanto, la utilidad del conjunto de entrenamiento no depende únicamente del número de imágenes, sino también de su representatividad, diversidad y calidad visual.

Este escenario motiva la adopción de un enfoque Data-Centric AI, centrado en mejorar el conjunto de datos antes que en modificar sucesivamente la arquitectura del clasificador. Para ello, este trabajo desarrolla un flujo de procesamiento que homogeneiza, amplía y refina los datos clínicos disponibles a partir de imágenes clave anotadas por especialistas y vídeos completos de colonoscopia.

1.1. Objetivos

1.1.1. Objetivo principal

El objetivo principal de este trabajo es evaluar si la mejora sistemática del conjunto de entrenamiento permite incrementar el rendimiento de la clasificación multiclase de pólipos colorectales sin modificar la arquitectura de los modelos. Para ello, se desarrolla un flujo de procesamiento que parte de imágenes clave anotadas por especialistas y de los vídeos de colonoscopia asociados, con el fin de homogeneizar, ampliar y refinar los datos disponibles.

La hipótesis de trabajo plantea que un conjunto de entrenamiento más representativo, diverso y menos redundante puede mejorar la capacidad de clasificación respecto al conjunto original. El flujo actúa exclusivamente sobre los datos de entrenamiento, mientras que los conjuntos de validación y test se mantienen fijos para permitir una comparación consistente entre las distintas versiones generadas.

La evaluación se realiza con EfficientNetB0 y ViT-Small como modelos de referencia. El rendimiento se analiza tras cada etapa del flujo mediante la F1-Score Macro, seleccionada por ser más informativa que la accuracy cuando existe desbalanceo entre clases[8]. De forma complementaria, se estudia el comportamiento por categoría para identificar mejoras o degradaciones específicas.

1.1.2. Objetivos específicos

- Ampliar el conjunto de entrenamiento a partir de los vídeos de colonoscopia asociados a las imágenes anotadas, extrayendo fotogramas cercanos al instante clínico de referencia mediante un proceso parametrizable de detección y seguimiento de lesiones.
- Refinar las muestras generadas para reducir la redundancia temporal y priorizar imágenes visualmente diversas y potencialmente más útiles para el entrenamiento.
- Comparar el conjunto original con las distintas versiones generadas del conjunto de entrenamiento, manteniendo fijos los modelos de referencia y los conjuntos de validación y test para analizar el efecto de cada etapa sobre la Macro-F1 y el rendimiento por categoría.

La Figura 1 resume el flujo de procesamiento aplicado sobre los datos clínicos.

2 ESTADO DEL ARTE

2.1. CNN y Transformers en imagen médica

En los últimos años, el análisis automático de imagen médica ha incorporado ampliamente modelos de aprendizaje profundo, especialmente redes neuronales convolucionales (*Convolutional Neural Networks*, CNN). Estas arquitecturas se han aplicado a tareas de clasificación, segmentación, localización y detección, gracias a su capacidad para aprender representaciones visuales directamente a partir de los datos[9].

Dentro de las CNN, EfficientNet propone un escalado conjunto de la profundidad, la anchura y la resolución de



Fig. 1: Flujo de procesamiento propuesto. A partir de imágenes anotadas y vídeos de colonoscopia, el conjunto de entrenamiento se normaliza, amplía mediante la selección de fotogramas próximos a las lesiones y se refina.

entrada, con el objetivo de mejorar la relación entre rendimiento y coste computacional[10]. Esta característica resulta adecuada para el presente trabajo, ya que permite utilizar un modelo de referencia con capacidad suficiente sin que el coste de los entrenamientos repetidos condicione la evaluación de las distintas versiones del conjunto de datos.

Más recientemente, los *Vision Transformers* (ViT) se han incorporado como una alternativa a las arquitecturas convolucionales en imagen médica. A diferencia de las CNN, procesan la imagen como una secuencia de regiones y utilizan mecanismos de atención para modelar relaciones entre distintas zonas visuales[11]. Las comparaciones recientes entre ambas familias indican que los Transformers presentan un potencial creciente, especialmente cuando se dispone de estrategias adecuadas de preentrenamiento y suficientes datos, aunque no sustituyen de forma directa a las CNN en todos los escenarios[12].

2.2. Clasificación mediante modelos de IA en colonoscopia

En el contexto de la colonoscopia, los sistemas de diagnóstico asistido por ordenador (*Computer-Aided Diagnosis*, CADx) se han orientado a estimar la histología de los pólipos a partir de imágenes endoscópicas. Algunos trabajos combinan además la localización y la caracterización de las lesiones; por ejemplo, PolyDeep se propone como un sistema capaz de detectar pólipos y clasificarlos en imágenes estáticas de colonoscopia[7].

Una parte importante de los trabajos existentes formula esta tarea como un problema de clasificación binaria o agrupada. Por ejemplo, Min et al. proponen un sistema CAD basado en imágenes de colonoscopia con *Linked Color Imaging* para diferenciar pólipos adenomatosos y no adenomatosos, mostrando un rendimiento comparable al de endoscopistas expertos en su conjunto de evaluación[6]. De forma similar, PolyDeep aborda la caracterización de pólipos distinguiendo lesiones neoplásicas y no neoplásicas[7]. Estos enfoques simplifican la diversidad histológica de los pólipos al agrupar varias tipologías en clases más amplias.

Frente a esta formulación binaria, algunos trabajos recientes han abordado la clasificación multiclase de pólipos directamente sobre imágenes de colonoscopia. Liu et al. proponen Polyp-DedNet, una red basada en ResNet50 modificada para tareas de clasificación de cuatro y seis clases[13]. Este tipo de trabajo es especialmente relevante para el presente proyecto, ya que evidencia que la clasificación histológica multiclase en imagen endoscópica es un problema activo.

No obstante, estos trabajos se centran principalmente en mejorar el rendimiento a partir del diseño del modelo, la

elección de la arquitectura o el uso de estrategias de entrenamiento específicas. En cambio, el presente trabajo aborda el problema desde una perspectiva complementaria: estudiar si la mejora sistemática del conjunto de entrenamiento puede incrementar el rendimiento del clasificador manteniendo constante la arquitectura del modelo.

2.3. Data-Centric AI

El paradigma *Data-Centric AI* plantea un cambio de enfoque respecto al desarrollo tradicional de sistemas de inteligencia artificial. En lugar de centrar la mejora del rendimiento únicamente en el diseño de nuevas arquitecturas o en la optimización del modelo, este enfoque propone actuar de forma sistemática sobre los datos utilizados para entrenar, evaluar y mantener el sistema. Zha et al. describen este paradigma como una disciplina orientada a la ingeniería de datos para IA, destacando que la calidad, cantidad, representatividad y mantenimiento de los datos son factores determinantes para el comportamiento final del modelo. Dentro de esta perspectiva, el desarrollo del conjunto de entrenamiento adquiere un papel central[14].

Esta idea se ha validado experimentalmente en trabajos como el de Bhatt et al., donde se compara un enfoque *model-centric* con uno *data-centric* manteniendo fija una arquitectura ResNet-18. En dicho estudio, la mejora del rendimiento no se obtiene mediante el diseño de un nuevo modelo, sino a través de intervenciones sobre el conjunto de datos, como el aumento de datos, la eliminación de duplicados mediante *multi-stage hashing* y la detección de etiquetas ruidosas mediante *confident learning*[15]. Aunque el contexto experimental de este trabajo se basa en datasets generales y no en imagen endoscópica, resulta relevante porque muestra que este tipo de técnicas pueden integrarse dentro de un flujo *data-centric* manteniendo constante la arquitectura del modelo.

En el presente trabajo, esta idea se adapta a un escenario de vídeo de colonoscopia, donde la redundancia no aparece únicamente como duplicación exacta, sino también como secuencias de frames consecutivos altamente similares. Por ello, la fase de reducción de redundancia temporal se plantea como parte del refinamiento del conjunto de entrenamiento.

2.4. Técnicas clásicas de visión artificial para el refinamiento de imágenes

Aunque los modelos de aprendizaje profundo constituyen la base de los sistemas actuales de clasificación, las técnicas clásicas de visión artificial siguen siendo útiles en fases previas de preparación y refinamiento de datos. En vídeo endoscópico, la calidad de los frames puede verse

afectada por problemas propios de la exploración, como variaciones de iluminación, desenfoque, reflejos, oclusiones, contacto con la mucosa o alta redundancia temporal entre imágenes consecutivas. Las revisiones sobre procesamiento de vídeo endoscópico destacan precisamente la importancia de aplicar técnicas de análisis visual y preprocesamiento antes de utilizar las imágenes en tareas posteriores de detección, clasificación o recuperación de información[16].

En este contexto, las técnicas clásicas permiten definir criterios interpretables y computacionalmente eficientes para descartar imágenes poco informativas o seleccionar muestras más representativas. Para la reducción de redundancia temporal pueden utilizarse medidas de similitud visual como Perceptual Hashing y Structural Similarity Index Measure (SSIM), que permiten comparar frames y detectar imágenes altamente similares desde perspectivas perceptuales y estructurales[17, 18]. Para la estimación de nitidez, las medidas de foco basadas en operadores diferenciales, como la Varianza del Laplaciano, proporcionan una aproximación clásica al grado de enfoque de una imagen[19].

3 CONJUNTO DE DATOS

Para el desarrollo del trabajo se utilizaron dos fuentes de datos. La primera corresponde al conjunto principal, proporcionado por el Hospital Clínic de Barcelona[20], e incluye imágenes anotadas y vídeos completos de colonoscopia obtenidos con equipos Olympus. A partir de estos datos se construyen los conjuntos de entrenamiento y validación, sobre los que se aplica y analiza el flujo de procesamiento.

El conjunto de test es externo e independiente, y se compone de una concatenación parcial de PIBAdb[21] y PICCOLO[22]. Estas fuentes proceden de otros hospitales y utilizan dispositivos de adquisición distintos, incluyendo equipos Olympus y Fujifilm. Por tanto, además de evaluar el rendimiento sobre imágenes no utilizadas durante el entrenamiento, el test representa un cambio de dominio respecto a los datos del Hospital Clínic, más próximo a un escenario mas general donde la prevalencia de las lesiones puede ser distinta y se opera con distintos fabricantes.

Como apoyo visual adicional, el Apéndice C presenta una proyección *t-SNE* de representaciones visuales de ambos conjuntos. Aunque parte de las muestras del test externo se solapa con los datos utilizados en entrenamiento, también aparecen agrupaciones externas en regiones periféricas de la proyección. Este patrón constituye una evidencia cualitativa de que algunas apariencias visuales del test están menos representadas en los datos del Hospital Clínic, lo que resulta coherente con el cambio de dominio introducido por los distintos centros y dispositivos de adquisición.

La Figura 2 muestra ejemplos representativos de las cuatro categorías histológicas en el conjunto principal y en el test externo.

El conjunto principal está compuesto por 3418 imágenes pertenecientes a cuatro clases: adenomas, adenomas serrados sésiles, pólipos hiperplásicos y adenocarcinomas. Sobre este conjunto se aplica un *split* estratificado de *train* y *validation*, asignando el 80 % de las muestras al conjunto de entrenamiento y el 20 % restante al conjunto de validación. Como se muestra en la Tabla 1, el conjunto de test contiene 1299 imágenes y mantiene las mismas categorías histológicas.

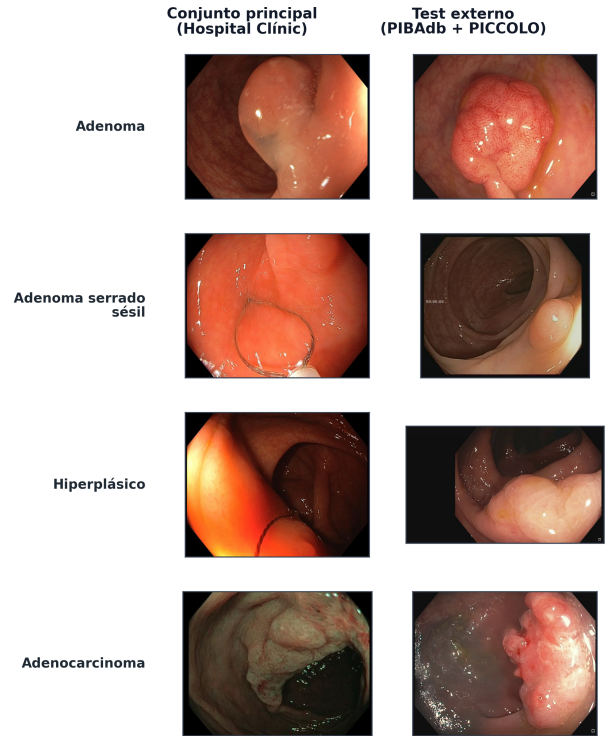


Fig. 2: Ejemplos representativos de las cuatro categorías histológicas en el conjunto principal del Hospital Clínic de Barcelona y en el conjunto de test externo, compuesto por imágenes procedentes de PIBAdb y PICCOLO.

Clase	Train	Validation	Test
Adenomas	1578	395	428
Adenomas serrados sésiles	725	182	214
Hiperplásicos	334	84	428
Adenocarcinomas	101	19	229
Total	2738	680	1299

TABLA 1: DISTRIBUCIÓN DE MUESTRAS POR CLASE EN LOS CONJUNTOS DE ENTRENAMIENTO, VALIDACIÓN Y TEST.

4 METODOLOGÍA

El desarrollo del proyecto se organizó siguiendo un enfoque ágil inspirado en Scrum. El trabajo se dividió en iteraciones cortas, centradas en implementar, validar y ajustar progresivamente los distintos componentes del flujo de procesamiento. Esta organización permitió adaptar la planificación ante incidencias técnicas, como cambios en los criterios de selección de imágenes, limitaciones de los detectores empleados o resultados experimentales no concluyentes.

El seguimiento se realizó mediante reuniones periódicas con el tutor, aproximadamente cada dos semanas, en las que se revisaron los avances, se validaron las decisiones técnicas y se priorizaron las tareas siguientes. Cuando fue necesario, se realizaron sesiones adicionales para concretar requisitos o analizar resultados específicos. Este proceso permitió mantener una evolución incremental del proyecto y ajustar las decisiones de desarrollo a la evidencia obtenida.

durante la experimentación.

5 DESARROLLO

Tras definir los objetivos y los conjuntos de datos utilizados, esta sección describe el desarrollo técnico del proyecto. El flujo de procesamiento consta de una fase inicial de normalización, seguida de tres etapas: definición de los modelos de referencia, ampliación mediante vídeo y refinamiento del conjunto de datos. La configuración de cada etapa se seleccionó empíricamente a partir de los experimentos realizados, priorizando aquella que obtuvo el mejor rendimiento en el conjunto de validación.

5.1. Fase 0: Normalización del conjunto de datos

Antes de realizar el *train-validation split* se aplica una etapa previa de normalización del dataset, de forma que todas las fases del pipeline trabajan sobre imágenes con el mismo preprocesado. Del mismo modo, el conjunto de test externo se procesa con la misma lógica de preprocesado.

Esta normalización consiste en eliminar las regiones negras presentes en las imágenes de colonoscopia, generadas a causa de que el área endoscópica no ocupa toda la imagen almacenada. Estas zonas no contienen información relevante para la clasificación y aumentan el área procesada por el modelo sin aportar contenido útil. La eliminación de la máscara negra periférica es una operación habitual en el preprocesado de imágenes de colonoscopia, pudiendo abordarse mediante umbralización y recorte de la región útil[23].

Dicho recorte se realiza de manera automática. Cada imagen se convierte a escala de grises y se umbraliza con un valor fijo de intensidad igual a 15, con el objetivo de identificar píxeles muy próximos al negro asociados a la zona periférica. A partir de la máscara binaria resultante se extraen los contornos externos y se selecciona el de mayor área como aproximación al campo endoscópico útil. Sobre esta región se calcula una área delimitadora, ampliada con un margen de 10 píxeles para evitar recortes excesivamente ajustados. Además de generar la imagen recortada, se almacenan las coordenadas, permitiendo así conservar la correspondencia espacial entre la imagen original y la región utilizada, información relevante para próximas fases del pipeline.

Una vez normalizadas las imágenes, el conjunto principal se divide en entrenamiento y validación mediante un *split* estratificado por categoría histológica, manteniendo la distribución de clases mostrada en la Tabla 1. El 80 % de las muestras se asigna al conjunto de entrenamiento y el 20 % restante al de validación.

La división se realiza agrupando las imágenes por identificador de lesión. De este modo, todas las imágenes asociadas a un mismo pólipo se asignan a una única partición. Esta restricción evita que distintas imágenes de la misma lesión aparezcan simultáneamente en entrenamiento y validación, lo que introduciría *Data leakage* y produciría una estimación excesivamente optimista del rendimiento.

5.2. Definición de los modelos de referencia

La primera fase del pipeline consiste en configurar los modelos de referencia que actúan como base experimental durante el proyecto. En un enfoque *Data-Centric AI* como el aplicado en este trabajo, esta decisión es fundamental, ya que el objetivo no es modificar las arquitecturas en cada fase, sino mantener constantes los modelos y su configuración de entrenamiento para poder atribuir la variaciones de rendimiento a los cambios introducidos en el conjunto de datos.

Por este motivo, en esta fase se definen las arquitecturas utilizadas, la receta de entrenamiento y la construcción del conjunto de validación.

5.2.1. Arquitecturas evaluadas

Como modelos de referencia se utilizan dos arquitecturas preentrenadas sobre ImageNet[24]: EfficientNetB0 y ViT-Small, estas arquitecturas permiten evaluar el comportamiento del pipeline sobre dos familias de modelos distintas: una red neuronal convolucional y un *Vision Transformer*. En ambos casos, el clasificador original se sustituye adaptándose a las cuatro clases sobre las que se trabajan: adenomas, adenomas serrados sésiles, pólipos hiperplásicos y adenocarcinomas.

Como red neuronal convolucional (CNN) se utiliza EfficientNetB0. Esta arquitectura se basa en ajustar de forma coordinada la profundidad, la anchura y la resolución de entrada de la red, lo que proporciona una relación favorable entre rendimiento y coste computacional[10]. Dentro de la familia EfficientNet, la versión B0 corresponde al modelo base, siendo la escogida para este proyecto por ofrecer una capacidad suficiente sin introducir un coste excesivo en los entrenamientos repetidos necesarios para evaluar las fases del pipeline.

Como *Vision Transformer* se utiliza ViT-Small. A diferencia de las CNN, que extraen patrones visuales mediante operaciones convolucionales sobre la imagen completa, los Vision Transformers dividen la imagen en regiones, proyectando cada una de ellas a una representación vectorial y procesan la secuencia resultante mediante mecanismos de atención[25]. Esto permite modelar relaciones globales entre distintas zonas de la imagen, aportando una perspectiva complementaria. La versión Small se selecciona con el mismo fin que en el caso de B0 en EfficientNet, una capacidad suficiente a un coste computacional y temporal moderado para la repetición de los experimentos.

5.2.2. Configuración de entrenamiento

Una vez definidas las arquitecturas de referencia, se fija una configuración de entrenamiento común para los experimentos del proyecto. La búsqueda inicial se realiza mediante *bayesian search*[26] sobre el conjunto de entrenamiento generado tras la Fase 0, utilizando EfficientNetB0 como arquitectura de calibración.

Con el objetivo de asegurar la estabilidad de los resultados frente a factores aleatorios (inicialización, data augmentation,...), y manteniendo un coste computacional viable, cada configuración se evalúa teniendo en cuenta tres ejecuciones con semillas distintas. La receta final se selecciona a partir de la media de la F1-Score Macro en valida-

ción y, en caso de empate, se utiliza el coeficiente de correlación de Matthews (MCC) como criterio secundario[27]. Una vez seleccionada, esta configuración se aplica tanto a EfficientNetB0 como a ViT-Small y se mantiene fija en las fases posteriores. ViT-Small no interviene en la selección de la receta ni en las decisiones de cada fase, sino que permite comprobar si el efecto de los cambios en el conjunto de entrenamiento se mantiene en una familia de modelos distinta.

La receta final incluye los parámetros generales de entrada, la estrategia de *fine-tuning*, la configuración del optimizador, mecanismos para tratar el desbalanceo, técnicas de regularización y los criterios de parada y selección del modelo. La Tabla A.1 resume la configuración utilizada en los experimentos.

Parámetros generales Los primeros parámetros de la receta vienen determinados por las arquitecturas evaluadas y por el uso de pesos preentrenados. El tamaño de entrada se fija en 224×224 píxeles y las imágenes se normalizan bajo los mismos criterios que en ImageNet, manteniendo así la coherencia con el preentrenamiento utilizado en EfficientNetB0 y ViT-Small[10, 24, 25]. El clasificador de cada arquitectura se adapta a las cuatro clases del problema: adenomas, adenomas serrados sésiles, pólipos hiperplásicos y adenocarcinomas.

El *batch size* se fija en 32 asegurando así la estabilidad del entrenamiento respecto a la capacidad computacional disponible. El número máximo de épocas se establece en 100, pero la finalización efectiva del entrenamiento queda regulada por el criterio de *early stopping*. Por ello, se define un límite superior suficiente para que se llegue al punto máximo de aprendizaje de la red.

Estrategia de fine-tuning El *fine-tuning* es una estrategia de *transfer learning* en la que el modelo no se entrena desde pesos aleatorios, sino a partir de una red previamente entrenada. En este trabajo se parte de modelos preentrenados sobre ImageNet, lo que permite aprovechar representaciones visuales generales y adaptarlas posteriormente al dominio de imágenes de colonoscopia.

La estrategia se divide en dos fases. En la primera, se congela el *backbone*, es decir, la parte de la red encargada de extraer representaciones visuales, y se entrena únicamente el clasificador, adaptado a las cuatro clases del problema. Esta fase inicial dura 3 épocas y utiliza un *learning rate* de 1×10^{-3} para la cabeza, permitiendo que la nueva capa final aprenda una primera separación entre clases sin modificar todavía los pesos preentrenados.

En la segunda fase, se descongela el conjunto completo de la red y se realiza el ajuste fino de todos los parámetros. Para ello, se utilizan tasas de aprendizaje diferenciadas: la cabeza clasificadora se actualiza con un *learning rate* de 1×10^{-4} , mientras que el *backbone* se actualiza con un valor menor, 1×10^{-5} . Esta diferencia permite adaptar progresivamente las representaciones preentrenadas al dominio de colonoscopia, evitando modificaciones demasiado bruscas en las capas encargadas de la extracción de características.

El uso de características preentrenadas y su posterior adaptación al nuevo dominio se apoya en la idea de que las representaciones aprendidas por una red pueden transferirse

a nuevas tareas, aunque su utilidad dependa de la relación entre la tarea original y la tarea destino [28].

Optimizador y evolución de learning rates El entrenamiento de los modelos se realiza mediante el optimizador AdamW, encargado de actualizar los pesos de la red durante el aprendizaje. AdamW es una variante de Adam que incorpora de forma adecuada regularización mediante *weight decay*[29]. El *weight decay* se emplea para penalizar pesos excesivamente grandes y reducir el riesgo de que el modelo se ajuste demasiado al conjunto de entrenamiento. En la receta final, este valor se fija en 5×10^{-4} .

Durante el entrenamiento se emplean *learning rates* diferenciados para el clasificador y el *backbone*, tal y como se ha descrito en el apartado anterior. Además, se utiliza el scheduler ReduceLROnPlateau, que reduce el *learning rate* cuando la métrica monitorizada deja de mejorar[30]. En este trabajo, el scheduler utiliza una *patience* de 4 épocas y un factor de reducción de 0.5, de modo que, si no se observa mejora durante 4 épocas consecutivas, el *learning rate* se reduce a la mitad para permitir ajustes más finos. También se define un valor mínimo de aprendizaje de 1×10^{-7} , evitando que la tasa de aprendizaje disminuya indefinidamente.

Finalmente, se aplica *gradient clipping* con valor 1.0 como medida de estabilidad durante el entrenamiento. Este mecanismo limita actualizaciones demasiado grandes de los pesos cuando los gradientes toman valores elevados, reduciendo el riesgo de inestabilidad durante el ajuste de los modelos.

Tratamiento del desbalanceo Dado el fuerte desbalanceo existente entre clases, es imprescindible combinar mecanismos a nivel de función de pérdida y a nivel de muestreo. El objetivo es aumentar la importancia de las clases minoritarias durante el entrenamiento sin modificar directamente las imágenes ni las etiquetas del dataset.

En primer lugar, se utiliza *Focal Loss* como función de pérdida. Esta función reduce el peso de los ejemplos que el modelo clasifica correctamente con facilidad y concentra más el aprendizaje en los casos difíciles. Esta propiedad resulta útil en datasets desbalanceados, ya que evita que el entrenamiento quede dominado por ejemplos sencillos o por clases mayoritarias.

Además, la función de pérdida incorpora pesos de clase calculados a partir de la inversa de la frecuencia de cada categoría. De forma conceptual, el peso de cada clase se calcula como $w_c = (1/n_c)^\alpha$, donde n_c es el número de muestras de la clase y α es el exponente de ponderación. En este caso se utiliza $\alpha = 0,7$, aplicando una corrección parcial del desbalanceo.

En segundo lugar, se utiliza un muestreador ponderado, el *WeightedRandomSampler*, durante la construcción de los *batches*. Este mecanismo modifica la probabilidad con la que cada muestra es seleccionada durante el entrenamiento, aumentando la presencia de clases minoritarias. De este modo, el tratamiento del desbalanceo se reparte entre la función de pérdida y el muestreo, buscando corregir el sesgo hacia las clases mayoritarias sin incrementar demasiado el riesgo de sobreajuste por repetición de ejemplos.

Regularización y early stopping Para reducir el riesgo de *overfitting* se incorporan varios mecanismos de regularización durante el entrenamiento. En primer lugar, se utiliza *Dropout* en el clasificador, este desactiva aleatoriamente parte de las activaciones durante el entrenamiento. De este modo, se evita que el modelo dependa siempre de las mismas combinaciones internas de características y se favorece una mayor capacidad de generalización[31].

Además, se emplea *stochastic depth*, también conocido como *drop path*, este desactiva aleatoriamente algunas rutas internas del modelo durante el entrenamiento, lo que obliga a la red a no depender de un único camino de procesamiento. Su objetivo es mejorar la robustez del modelo y reducir el sobreajuste.

La función de *loss* incorpora también *label smoothing*. Este mecanismo suaviza ligeramente las predicciones, evitando que el modelo aprenda las clases como certezas absolutas. En lugar de forzar una probabilidad de 1 para la clase correcta y 0 para el resto, se reparte una pequeña parte de la probabilidad entre las demás clases. Esto reduce el exceso de confianza del modelo y puede mejorar su capacidad de generalización.

Finalmente, se utiliza *early stopping* con una *patience* de 16 épocas. Este criterio detiene el entrenamiento cuando el rendimiento en validación deja de mejorar durante un número determinado de épocas consecutivas, evitando continuar el entrenamiento cuando ya no aporta una mejora real sobre datos no vistos[32]. El checkpoint final se selecciona directamente mediante la F1-Score Macro en validación.

Data augmentation Además de los hiperparámetros de entrenamiento, se define un *data augmentation* aplicado únicamente sobre el conjunto de entrenamiento. El objetivo de estas transformaciones es introducir variabilidad controlada en las imágenes, reduciendo el riesgo de que el modelo memorice patrones demasiado específicos y favoreciendo una mejor generalización[33]. En cambio, sobre los conjuntos de validación y test no se aplica aumento de datos, sino un preprocesado determinista formado por redimensionado, conversión a tensor y normalización como ImageNet.

El aumento de datos realiza transformaciones compatibles con la variabilidad esperable en imágenes endoscópicas. En primer lugar, se aplica *RandomResizedCrop*, que genera recortes aleatorios de la imagen modificando ligeramente la escala y la proporción, simulando diferencias de encuadre y distancia respecto a la lesión. También se utilizan giros verticales y horizontales, así como una transformación moderada que introduce pequeñas rotaciones, traslaciones y cambios de escala. Estas operaciones aumentan la diversidad espacial del conjunto de entrenamiento.

Además, se aplica *ColorJitter* con variaciones de brillo, contraste, saturación y tono, con el objetivo de reflejar cambios de iluminación y adquisición propios de las imágenes de colonoscopia. Finalmente, se incorporan *GaussianBlur* y *RandomErasing* como mecanismos adicionales de regularización visual. El primero introduce desenfoques ligeros de forma aleatoria, mientras que el segundo oculta pequeñas regiones de la imagen durante el entrenamiento, obligando al modelo a no depender de una única zona visual concreta. La Tabla A.2 resume las transformaciones utilizadas.

5.3. Ampliación mediante vídeo

La segunda fase del flujo de procesamiento tiene como objetivo ampliar el conjunto de entrenamiento a partir de los vídeos completos de las exploraciones, asociados a las imágenes etiquetadas. Cada imagen original proporciona una clase histológica y un instante aproximado dentro del vídeo en el que aparece la lesión. A partir de esta información, se pueden extraer nuevos frames cercanos al instante temporal de referencia con el objetivo de incorporar muestras adicionales.

Esta fase se aplica únicamente sobre el conjunto de entrenamiento, manteniendo sin modificar los conjuntos de validación y test. De este modo, las evaluaciones posteriores permiten analizar el efecto de los cambios introducidos en los datos de entrenamiento sin alterar los conjuntos utilizados para comparar el rendimiento de los modelos.

En esta fase se plantean dos estrategias distintas de ampliación. Ambas comparten el mismo proceso inicial, pero difieren en el criterio utilizado para decidir qué frames se incorporan finalmente al conjunto de entrenamiento. La primera estrategia está orientada a compensar parcialmente el desbalanceo entre clases mediante una selección dependiente de la histología, mientras que la segunda aplica un criterio más restrictivo basado en la confianza del detector. En las siguientes subsecciones se describe primero el proceso común y, posteriormente, las diferencias metodológicas entre ambas estrategias.

5.3.1. Proceso común de extracción de frames

El proceso de ampliación parte de las imágenes etiquetadas del conjunto de entrenamiento. Para cada muestra se utiliza la información proporcionada para localizar el vídeo de colonoscopia correspondiente y calcular el instante temporal asociado a la imagen original. Este instante se utiliza como referencia para analizar únicamente una ventana temporal acotada de ± 3 segundos, evitando procesar el vídeo completo y reduciendo el coste computacional de la fase.

Dentro de esta ventana, los frames se muestrean a una frecuencia de 5 FPS. Esta frecuencia permite obtener varias muestras cercanas al instante clínico de referencia sin introducir todos los frames consecutivos del vídeo, que suelen presentar una alta redundancia temporal. Antes de aplicar la detección, cada frame se recorta utilizando la misma región clínica útil definida en la fase de normalización. Cuando las coordenadas de recorte de la imagen original están disponibles, estas se reutilizan sobre el frame de vídeo para mantener la coherencia espacial entre la imagen etiquetada y las nuevas muestras generadas. Si no existen coordenadas previas, el recorte se calcula automáticamente siguiendo el mismo procedimiento descrito en la Fase 0.

Sobre los frames preprocesados se aplica un detector YOLOv8 como componente auxiliar para localizar posibles regiones de pólipo[34, 35]. Este detector no ha sido entrenado como parte del proyecto, sino que procede de un modelo entrenado sobre CVC-ClinicDB[36] y descargado del repositorio asociado a YOLO-SAM2[37]. Por tanto, su función dentro del pipeline no es resolver la clasificación histológica, sino proporcionar regiones candidatas que permitan guiar la selección de frames.

Las detecciones obtenidas se combinan con ByteTrack, siguiendo una estrategia de *tracking-by-detection*: primero

se detectan posibles pólipos en cada frame y después se asocian las detecciones entre frames cercanos para construir trayectorias temporales[38]. Esta asociación temporal permite reducir la dependencia de detecciones aisladas, ya que una región que aparece de forma persistente en varios frames resulta más fiable que una detección puntual.

Dentro de cada ventana temporal pueden aparecer varias trayectorias candidatas. Por este motivo, se selecciona una trayectoria principal priorizando, en primer lugar, aquella con mayor persistencia temporal; en segundo lugar, la más cercana al timestamp original; y, finalmente, la de mayor confianza media de detección. A partir de esta trayectoria principal, cada estrategia de ampliación aplica su propio criterio para decidir qué frames se incorporan finalmente al conjunto de entrenamiento.

5.3.2. Estrategia de balanceo por histología

La primera estrategia de ampliación se diseña con el objetivo de ampliar el conjunto de entrenamiento teniendo en cuenta el desbalanceo existente entre clases. En lugar de extraer el mismo número de frames para todas las histologías, el número máximo de candidatos generados por cada imagen se ajusta en función de la frecuencia de su clase en el conjunto de entrenamiento.

Para ello, se toma como referencia la frecuencia de la clase mayoritaria y se calcula, para cada histología, una cantidad máxima de candidatos inversamente proporcional a su frecuencia. De este modo, las clases más representadas generan menos frames adicionales, mientras que las clases minoritarias pueden incorporar un mayor número de muestras. Esta cantidad se limita mediante un máximo global de candidatos por vídeo, evitando que una misma exploración genere un número excesivo de imágenes muy similares.

Esta estrategia permite que la fase de ampliación no solo aumente el tamaño del dataset, sino que también contribuya parcialmente a corregir el desbalanceo desde la propia generación de datos. La intención no es obtener un conjunto completamente equilibrado, sino reducir la diferencia entre clases sin modificar manualmente las etiquetas ni introducir datos externos.

En esta variante, si el detector no encuentra una trayectoria válida dentro de la ventana temporal, se aplica una estrategia alternativa basada en cercanía temporal. En ese caso, se seleccionan los frames muestreados más próximos al timestamp original. Esta decisión se basa en que el instante temporal procede de una imagen previamente etiquetada por especialistas, por lo que los frames cercanos pueden seguir conteniendo información útil sobre la misma lesión aunque el detector no haya producido una trayectoria fiable.

5.3.3. Estrategia de ampliación filtrada por confianza

La segunda estrategia de ampliación modifica el criterio de selección utilizado en la variante anterior. En este caso, el objetivo no es limitar el número de frames en función de la histología, sino aceptar únicamente aquellos candidatos que presentan una confianza mínima de detección. Por tanto, la restricción principal no se aplica sobre la cantidad de imágenes generadas por cada clase, sino sobre la fiabilidad individual de cada frame seleccionado.

Para ello, se establece un umbral mínimo de confianza de detección. En esta configuración, solo se consideran válidos

los frames cuya detección asociada alcanza una confianza igual o superior a 0,60. De este modo, se evita incorporar imágenes en las que el detector localiza la región candidata con baja seguridad, reduciendo la probabilidad de añadir frames mal localizados o con una evidencia débil de presencia del pólipo.

A diferencia de la estrategia balanceada por histología, esta variante no utiliza un número máximo de candidatos dependiente de la frecuencia de cada clase. En su lugar, todas las histologías comparten el mismo límite superior global de candidatos por vídeo, definido para evitar que una misma exploración añada una cantidad excesiva de frames. Como consecuencia, aunque el criterio de aceptación de cada frame es más exigente, el número total de imágenes generadas puede ser mayor que en la estrategia balanceada, especialmente cuando el detector mantiene una confianza alta en varios frames de la ventana temporal.

Otra diferencia relevante es que esta estrategia no incorpora frames únicamente por cercanía temporal cuando no existe una detección suficientemente fiable. Si dentro de la ventana analizada no se obtiene una trayectoria válida que cumpla el umbral de confianza, no se añaden nuevos frames para esa muestra. De este modo, la estrategia prioriza que los frames incorporados estén respaldados por una detección fiable, mientras que la estrategia balanceada prioriza una generación más controlada en función del desbalanceo entre clases.

Esta segunda variante permite comparar dos formas distintas de utilizar los vídeos dentro del pipeline: una orientada a regular la cantidad de imágenes añadidas según la histología y otra orientada a filtrar los candidatos según la confianza del detector. Ambas estrategias generan conjuntos de entrenamiento distintos que posteriormente se utilizan como entrada para la fase de refinamiento y la evaluación experimental.

Clase	Train original	Ampliación balanceada	Ampliación por confianza
Adenomas	1578	2850	8313
Adenomas serrados sésiles	725	2554	3047
Hiperplásicos	334	1613	1421
Adenocarcinomas	101	961	515
Total	2738	7978	13296

TABLA 2: DISTRIBUCIÓN DE IMÁGENES POR CLASE TRAS LAS ESTRATEGIAS DE AMPLIACIÓN DE VÍDEO.

Como se observa en la Tabla 2, ambas estrategias incrementan el tamaño del conjunto de entrenamiento respecto al conjunto original, pero lo hacen siguiendo criterios diferentes.

5.4. Refinamiento del conjunto de datos

La tercera fase del pipeline se orienta al refinamiento de los conjuntos de entrenamiento generados tras la ampliación mediante vídeo. En un enfoque Data-Centric AI, aumentar el número de imágenes disponibles no garantiza por sí mismo una mejora del clasificador, ya que las nuevas muestras pueden introducir redundancia temporal o aportar poco valor informativo. Por ello, esta fase busca reducir la redundancia y conservar las muestras más útiles de cada conjunto generado en la fase anterior[14, 39, 15].

El refinamiento se aplica exclusivamente sobre el conjunto de entrenamiento y parte de las dos variantes generadas

en la fase anterior. En la variante de ampliación balanceada, se eliminan previamente los frames con una confianza de YOLO inferior a 0,4. Sobre ambas variantes se evaluaron estrategias guiadas por el modelo, filtros clásicos de calidad visual y reducción de redundancia temporal. La configuración final se seleccionó empíricamente según su rendimiento en validación y se centra en la reducción de redundancia temporal, al ser la estrategia que proporcionó el mejor resultado.

5.4.1. Reducción de redundancia temporal

La reducción de redundancia temporal busca evitar que los frames extraídos de vídeo aumenten el tamaño del dataset sin aportar diversidad visual real. Para ello, se identifican grupos de imágenes altamente similares dentro de un mismo contexto clínico y se conservan únicamente las muestras más útiles de cada grupo.

Para evitar comparaciones entre imágenes de lesiones distintas, las muestras se agrupan primero por identificador de lesión. Dentro de cada grupo, las imágenes se ordenan temporalmente mediante el campo *elapsed seconds* y se dividen en eventos temporales próximos. De este modo, la comparación de similitud se restringe a imágenes correspondientes al mismo pólipo y a instantes cercanos dentro de la exploración.

Dentro de cada evento temporal se comparan todos los pares posibles de imágenes mediante dos medidas complementarias: *Structural Similarity Index Measure* (SSIM) y *Perceptual Hashing* (pHash)[18, 17]. La primera permite comparar la similitud estructural entre dos imágenes, mientras que la segunda proporciona una representación que permite detectar imágenes visualmente similares. Un par de imágenes se considera redundante únicamente cuando cumple simultáneamente ambos criterios, es decir, cuando presenta una alta similitud estructural y una distancia perceptual reducida. Esta combinación hace que el criterio de descarte sea más conservador que usar una única métrica de similitud.

Los pares redundantes se representan posteriormente como un grafo, donde cada imagen corresponde a un nodo y cada relación de redundancia a una arista. Cada componente conectada del grafo se interpreta como un grupo de redundancia. Esto permite agrupar imágenes de manera que si una imagen es redundante con una segunda, y esta segunda lo es con una tercera, las tres quedan incluidas en el mismo grupo.

Una vez identificados los grupos redundantes, se seleccionan las mejores imágenes dentro de cada grupo mediante un *Quality Score*. Este score combina tres criterios normalizados dentro del propio grupo: la nitidez de la imagen, estimada mediante la varianza del Laplaciano (L); el área relativa de la detección (A), que aproxima la presencia visible de la lesión; y la confianza del detector (C). La expresión utilizada es:

$$Q = 0,30 \cdot L + 0,20 \cdot A + 0,50 \cdot C$$

Finalmente, la selección no conserva un número fijo de imágenes para todas las clases, sino que adapta el número de muestras retenidas en función de la histología. Las clases más frecuentes conservan menos imágenes dentro de cada

grupo redundante, mientras que las clases menos representadas pueden conservar más muestras, buscando reducir la redundancia temporal sin penalizar en exceso a las clases minoritarias.

5.4.2. Configuración final de la fase

Además de la reducción de redundancia temporal, se evaluaron de forma exploratoria una selección guiada por la confianza del clasificador de referencia y filtros clásicos de calidad visual. Estas alternativas no mejoraron el rendimiento en validación, por lo que no se incorporaron a la configuración final.

La configuración final de la Fase 3 se centra, por tanto, en la reducción de redundancia temporal. En la evaluación se consideran tres variantes principales del conjunto de entrenamiento: una variante de entrada con candidatos de vídeo filtrados por confianza mínima de 0,4, sin refinamiento adicional, y dos variantes con deduplicación temporal aplicada sobre los conjuntos generados en la Fase 2. Esta comparación permite analizar tanto el efecto del filtrado previo por confianza como el impacto específico de la deduplicación temporal sobre el rendimiento del clasificador.

La distribución final de imágenes retenidas en cada variante se detalla en la Tabla B.1.

6 RESULTADOS

En esta sección se presentan los resultados obtenidos durante la evaluación experimental del pipeline. Cada valor corresponde a la media y desviación estándar de la F1-Score Macro obtenida con cuatro semillas distintas. La Tabla 3 compara el rendimiento de las distintas variantes del conjunto de entrenamiento sobre los conjuntos de validación y test.

La Tabla 3 resume la Macro-F1 media y su desviación estándar sobre cuatro semillas para cada variante. En validación, todas las variantes generadas a partir de vídeo superan al conjunto original en ambas arquitecturas. En EfficientNetB0, la ampliación filtrada por confianza alcanza la mayor media ($0,586 \pm 0,010$ frente a $0,537 \pm 0,035$), mientras que su versión deduplicada conserva un resultado prácticamente equivalente con menor dispersión entre ejecuciones. En ViT-Small, la mayor media corresponde a esta última variante ($0,623 \pm 0,019$ frente a $0,595 \pm 0,010$), aunque la diferencia respecto a la ampliación por confianza sin deduplicación es reducida y presenta una variabilidad mayor.

En el test externo, las mejoras no se reproducen de forma homogénea. La deduplicación sobre la ampliación balanceada es la única variante que supera al baseline en las cuatro combinaciones de arquitectura y conjunto de evaluación; en test, los incrementos son moderados, de 0,491 a 0,500 en EfficientNetB0 y de 0,482 a 0,491 en ViT-Small. Por el contrario, las variantes filtradas por confianza no mejoran de forma consistente el test y reducen el rendimiento de ViT-Small. Los resultados por clase se incluyen en las Tablas B.2 y B.3.

Variante	EfficientNetB0		ViT-Small	
	Validación	Test	Validación	Test
<i>Referencia</i>				
Conjunto original (<i>baseline</i>)	0,537 $\pm$ 0,035	0,491 $\pm$ 0,006	0,595 $\pm$ 0,010	0,482 $\pm$ 0,012
<i>Ampliación mediante vídeo</i>				
Balanceada	0,567 $\pm$ 0,006	0,500 $\pm$ 0,008	0,620 $\pm$ 0,009	0,464 $\pm$ 0,018
Filtrada por confianza ($\text{conf}_{\text{YOLO}} \geq 0,60$)	0,586 $\pm$ 0,010	0,490 $\pm$ 0,006	0,618 $\pm$ 0,003	0,448 $\pm$ 0,015
<i>Refinamiento</i>				
Deduplicación sobre ampliación balanceada ($\text{conf}_{\text{YOLO}} > 0,4$)	0,571 $\pm$ 0,011	0,500 $\pm$ 0,014	0,616 $\pm$ 0,010	0,491 $\pm$ 0,007
Deduplicación sobre ampliación filtrada por confianza	0,583 $\pm$ 0,005	0,492 $\pm$ 0,010	0,623 $\pm$ 0,019	0,465 $\pm$ 0,018

TABLA 3: MACRO-F1 GLOBAL, EXPRESADA COMO MEDIA $\pm$ DESVIACIÓN ESTÁNDAR SOBRE CUATRO SEMILLAS.

7 DISCUSSION Y FUTURE WORK

Los resultados respaldan la hipótesis de partida. Ampliar el conjunto de entrenamiento desde vídeo y reducir su redundancia aporta mejoras dentro del dominio de entrenamiento, sin modificar las arquitecturas ni su receta de entrenamiento. La deduplicación temporal permite conservar el rendimiento con menos muestras redundantes, pero no muestra una ventaja sistemática sobre la ampliación en todas las configuraciones. Asimismo, el contraste entre la estrategia balanceada y la filtrada por confianza indica que el criterio basado en el detector puede ser útil en validación, pero es más dependiente del modelo y del dominio evaluado.

La falta de una mejora externa homogénea es coherente con el cambio de dominio entre el Hospital Clínic y las fuentes del test. La proyección *t-SNE* del Apéndice C muestra solapamiento parcial entre ambos conjuntos, junto con agrupaciones de muestras externas en regiones periféricas; aunque es una evidencia cualitativa, este patrón es compatible con apariencias visuales poco representadas durante el entrenamiento. Como continuación, sería necesario entrenar y validar el flujo con datos multicéntricos, incorporar distintos dispositivos de adquisición y adaptar el detector de la fase de ampliación al dominio objetivo.

8 CONCLUSIONES

Este trabajo ha evaluado si un flujo Data-Centric AI puede mejorar la clasificación multiclase de pólipos actuando sobre el conjunto de entrenamiento y manteniendo fijos los modelos de referencia. El flujo combina normalización, ampliación desde vídeo y reducción de redundancia temporal, y se evaluó con EfficientNetB0 y ViT-Small sobre particiones de validación y test fijas.

La hipótesis queda respaldada dentro del dominio de entrenamiento: las variantes generadas mejoran la Macro-F1 de validación en ambas arquitecturas. Sin embargo, la mejora no se traslada de forma consistente al test externo, en coherencia con el cambio de dominio observado cualitativamente en el Apéndice C. Por tanto, la principal aportación del trabajo es mostrar que la mejora sistemática de los datos puede beneficiar al aprendizaje dentro de su dominio de origen, mientras que la generalización a otros entornos requiere incorporar diversidad de centros y dispositivos.

9 USO DE IA

Durante el desarrollo de este proyecto se han utilizado herramientas de IA generativa como apoyo en dos tipos de tareas. En primer lugar, se ha utilizado para la revisión formal del texto, incluyendo aspectos de corrección lingüística y reformulación de párrafos ya redactados. En segundo lugar, se han utilizado como apoyo para comprender conceptos técnicos de visión por computador y deep learning no estudiados previamente.

La herramienta empleada con este fin ha sido Gemini 3.1 Pro[40].

A RECETA DE ENTRENAMIENTO

Parámetro	Valor
Arquitecturas	EfficientNetB0 / ViT-Small
Pesos iniciales	ImageNet
Número de clases	4
Tamaño de entrada	224×224
Normalización	ImageNet
Batch size	32
Max epochs	100
Estrategia de entrenamiento	<i>Fine-tuning</i> por fases
Warm-up	3 épocas
LR cabeza en warm-up	1×10^{-3}
LR cabeza en fine-tuning	1×10^{-4}
LR backbone en fine-tuning	1×10^{-5}
Optimizador	AdamW
Weight decay	5×10^{-4}
Scheduler	ReduceLROnPlateau
Scheduler patience	4
Scheduler factor	0.5
Minimum LR	1×10^{-7}
Gradient clipping	1.0
Loss	Focal Loss
Focal gamma	2.0
Class weights	Inversa de frecuencia, $\alpha = 0,7$
Sampler	WeightedRandomSampler
Sampler class exponent	0.4
Dropout	0.30
Stochastic depth / Drop path	0.10
Label smoothing	0.05
Early stopping patience	16
Selección de checkpoint	Macro-F1 en validación

TABLA A.1: RECETA DE ENTRENAMIENTO FINAL UTILIZADA EN LOS EXPERIMENTOS.

Transformación	Configuración
RandomResizedCrop	scale= (0.7, 1.0), ratio= (0.85, 1.18)
RandomVerticalFlip	$p = 0,5$
RandomHorizontalFlip	$p = 0,5$
RandomAffine	grados=20
Traslación	(0,08, 0,08)
Escala	(0,9, 1,1)
ColorJitter brightness	0,25
ColorJitter contrast	0,25
ColorJitter saturation	0,12
ColorJitter hue	0,03
GaussianBlur	$p = 0,2$, kernel=3, sigma=(0,1, 1,5)
RandomErasing	$p = 0,25$, scale=(0,02, 0,12)

TABLA A.2: TRANSFORMACIONES APLICADAS AL CONJUNTO DE ENTRENAMIENTO EN LA RECETA FINAL.

B RESULTADOS COMPLEMENTARIOS POR CLASE

Variante	Adenoma	SSA	Hiperplásico	Adenocarcinoma	Total
Conjunto original	1578	725	334	101	2738
Ampliación balanceada	2850	2554	1613	961	7978
Deduplicación sobre ampliación balanceada ($conf_{YOLO} > 0,4$)	2333	1782	1100	671	5886
Ampliación filtrada por confianza ($conf_{YOLO} \geq 0,60$)	8313	3047	1421	515	13296
Deduplicación sobre ampliación filtrada por confianza	5591	2613	1339	515	10058

TABLA B.1: NÚMERO DE IMÁGENES POR CLASE EN LAS DISTINTAS VARIANTES DEL CONJUNTO DE ENTRENAMIENTO.

Variante del conjunto	Validation				Test			
	Adenoma	SSA	Hiper.	Adenoc.	Adenoma	SSA	Hiper.	Adenoc.
Baseline	0,6174 ± 0,0367	0,4484 ± 0,0407	0,2477 ± 0,0057	0,8350 ± 0,0690	0,5660 ± 0,0097	0,4085 ± 0,0831	0,3292 ± 0,0852	0,6607 ± 0,0218
Ampliación balanceada	0,6409 ± 0,0107	0,4586 ± 0,0163	0,2511 ± 0,0233	0,9189 ± 0,0000	0,6023 ± 0,0147	0,4347 ± 0,0292	0,3093 ± 0,0181	0,6550 ± 0,0233
Deduplicación sobre ampliación balanceada ($conf_{YOLO} > 0,4$)	0,6495 ± 0,0151	0,4817 ± 0,0127	0,2362 ± 0,0350	0,9178 ± 0,0023	0,6032 ± 0,0073	0,4295 ± 0,0426	0,3046 ± 0,0191	0,6640 ± 0,0184
Ampliación filtrada por confianza ($conf_{YOLO} \geq 0,60$)	0,6350 ± 0,0080	0,5013 ± 0,0165	0,2839 ± 0,0286	0,9253 ± 0,0128	0,6019 ± 0,0072	0,4194 ± 0,0179	0,2855 ± 0,0354	0,6512 ± 0,0191
Deduplicación sobre ampliación filtrada por confianza	0,6419 ± 0,0221	0,4910 ± 0,0069	0,2799 ± 0,0140	0,9193 ± 0,0203	0,6010 ± 0,0069	0,4186 ± 0,0199	0,2903 ± 0,0293	0,6570 ± 0,0468

TABLA B.2: F1-SCORE POR CLASE PARA EFFICIENTNETB0 EN VALIDACIÓN Y TEST.

Variante del conjunto	Validation				Test			
	Adenoma	SSA	Hiper.	Adenoc.	Adenoma	SSA	Hiper.	Adenoc.
Baseline	0,6634 ± 0,0152	0,5556 ± 0,0037	0,2940 ± 0,0190	0,8668 ± 0,0250	0,6183 ± 0,0039	0,4056 ± 0,0250	0,2055 ± 0,0397	0,6972 ± 0,0170
Ampliación balanceada	0,6981 ± 0,0275	0,5591 ± 0,0237	0,3036 ± 0,0144	0,9176 ± 0,0226	0,6261 ± 0,0165	0,2821 ± 0,0435	0,2288 ± 0,0618	0,7176 ± 0,0115
Deduplicación sobre ampliación balanceada ($conf_{YOLO} > 0,4$)	0,6863 ± 0,0056	0,5560 ± 0,0116	0,2931 ± 0,0539	0,9287 ± 0,0312	0,6211 ± 0,0258	0,3030 ± 0,0320	0,3243 ± 0,0435	0,7154 ± 0,0290
Ampliación filtrada por confianza ($conf_{YOLO} \geq 0,60$)	0,7210 ± 0,0100	0,5893 ± 0,0141	0,3031 ± 0,0335	0,8582 ± 0,0462	0,6302 ± 0,0268	0,3493 ± 0,0238	0,1253 ± 0,0208	0,6854 ± 0,0352
Deduplicación sobre ampliación filtrada por confianza	0,7106 ± 0,0118	0,5765 ± 0,0091	0,3115 ± 0,0423	0,8945 ± 0,0371	0,6490 ± 0,0089	0,3438 ± 0,0310	0,1835 ± 0,0370	0,6849 ± 0,0548

TABLA B.3: F1-SCORE POR CLASE PARA ViT-SMALL EN VALIDACIÓN Y TEST.

C ANÁLISIS VISUAL DEL CAMBIO DE DOMINIO

La Figura C.1 muestra una proyección bidimensional de representaciones visuales extraídas mediante EfficientNet-B0 preentrenada en ImageNet.

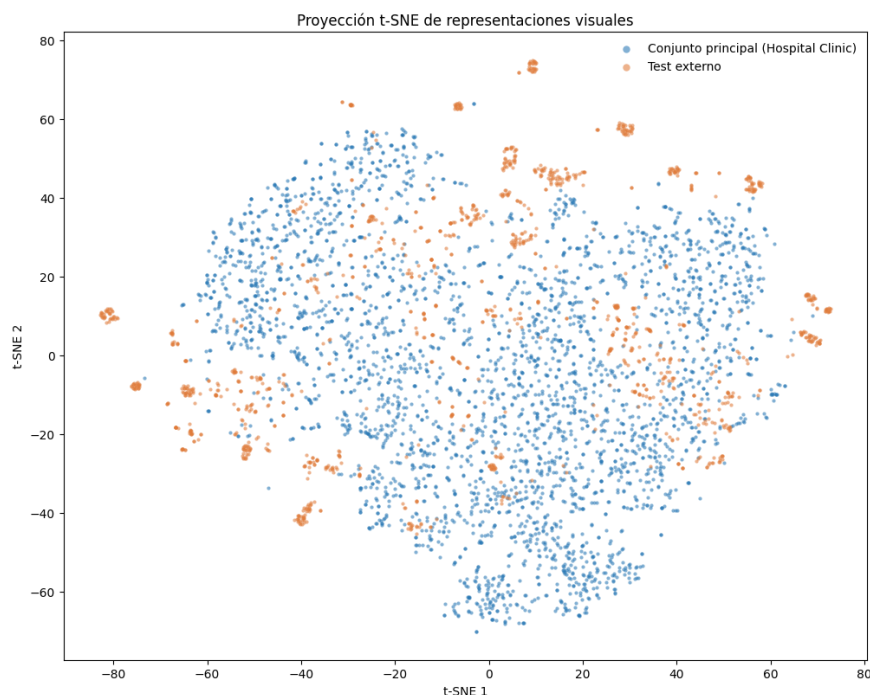


Fig. C.1: Proyección t -SNE de representaciones visuales del conjunto principal del Hospital Clínic y del test externo.

BIBLIOGRAFÍA

- [1] World Health Organization, “Cáncer colorrectal,” 2026.
- [2] American Cancer Society, “Pruebas para el diagnóstico del cáncer de colon o de recto,” 2024.
- [3] Mayo Clinic, “Colonoscopia,” 2026.
- [4] Mayo Clinic, “Pólipos en el colon,” 2026.
- [5] B. K. Abu Dayyeh, N. Thosani, V. Konda, M. B. Wallace, D. K. Rex, S. S. Chauhan, J. H. Hwang, S. Komanduri, M. Manfredi, J. T. Maple, F. M. Murad, U. D. Siddiqui, and S. Banerjee, “ASGE technology committee systematic review and meta-analysis assessing the ASGE PIVI thresholds for adopting real-time endoscopic assessment of the histology of diminutive colorectal polyps,” *Gastrointestinal Endoscopy*, vol. 81, no. 3, pp. 502.e1–502.e16, 2015.
- [6] M. Min, S. Su, W. He, Y. Bi, Z. Ma, and Y. Liu, “Computer-aided diagnosis of colorectal polyps using linked color imaging colonoscopy to predict histology,” *Scientific Reports*, vol. 9, p. 2881, 2019.
- [7] P. Davila-Piñón, A. Nogueira-Rodríguez, A. I. Díez-Martín, L. Codesido, J. Herrero, M. Puga, L. Rivas, E. Sánchez, F. Fdez-Riverola, D. Glez-Peña, M. Reboiro-Jato, H. López-Fernández, and J. Cubiella, “Optical diagnosis in still images of colorectal polyps: Comparison between expert endoscopists and PolyDeep, a computer-aided diagnosis system,” *Frontiers in Oncology*, vol. 14, p. 1393815, 2024.
- [8] M. Salmi, D. Atif, D. Oliva, A. Abraham, and S. Ventura, “Handling imbalanced medical datasets: review of a decade of research,” *Artificial Intelligence Review*, vol. 57, 2024.
- [9] D. R. Sarvamangala and R. V. Kulkarni, “Convolutional neural networks in medical image understanding: A survey,” *Evolutionary Intelligence*, vol. 15, no. 1, pp. 1–22, 2021.
- [10] M. Tan and Q. V. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks,” *arXiv preprint arXiv:1905.11946*, 2019.
- [11] F. Shamshad, S. Khan, S. W. Zamir, M. H. Khan, M. Hayat, F. S. Khan, and H. Fu, “Transformers in medical imaging: A survey,” *arXiv preprint arXiv:2201.09873*, 2022.
- [12] S. Takahashi, Y. Sakaguchi, N. Kouno, K. Takasawa, et al., “Comparison of vision transformers and convolutional neural networks in medical image analysis: A systematic review,” *Journal of Medical Systems*, vol. 48, no. 1, p. 84, 2024.
- [13] S. Liu, X. Liu, S. Chang, Y. Sun, K. Li, Y. Hou, Z. Qin, and X. Zhang, “Multi-classification of polyps in colonoscopy images based on an improved deep convolutional neural network,” *Computers, Materials & Continua*, vol. 75, no. 3, pp. 5837–5852, 2023.
- [14] D. Zha, Z. P. Bhat, K.-H. Lai, F. Yang, Z. Jiang, S. Zhong, and X. Hu, “Data-centric artificial intelligence: A survey,” *arXiv preprint arXiv:2303.10158*, 2023.
- [15] N. Bhatt, N. Bhatt, P. Prajapati, V. Sorathiya, S. Alshathri, and W. El-Shafai, “A data-centric approach to improve performance of deep learning models,” *Scientific Reports*, vol. 14, no. 1, p. 22329, 2024.
- [16] B. Münzer, K. Schoeffmann, and L. Böszörményi, “Content-based processing and analysis of endoscopic images and videos: A survey,” *Multimedia Tools and Applications*, vol. 77, pp. 1323–1362, 2018.
- [17] S. McKeown and W. J. Buchanan, “Hamming distributions of popular perceptual hashing techniques,” *Forensic Science International: Digital Investigation*, vol. 44, p. 301509, 2023.
- [18] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.
- [19] S. Pertuz, D. Puig, and M. A. Garcia, “Analysis of focus measure operators for shape-from-focus,” *Pattern Recognition*, vol. 46, no. 5, pp. 1415–1432, 2013.
- [20] Hospital Clínic Barcelona, “Hospital clínic barcelona.” <https://www.clinicbarcelona.org/>, s.f. Accedido el: 4 de junio de 2026.
- [21] A. Nogueira-Rodríguez, R. Dominguez-Carbajales, F. Campos-Tato, J. Herrero, M. Puga, L. Rivas, E. Sanchez, A. Iglesias, L. Codesido, A. I. Díez-Martín, P. Davila-Piñón, J. Cubiella, F. Fernandez-Riverola, H. Lopez-Fernandez, D. Glez-Pena, and M. Reboiro-Jato, “PIBAdb: a public cohort of multimodal colonoscopy videos and images including polyps with histological information,” *Computer Methods and Programs in Biomedicine*, vol. 280, p. 109315, 2026.
- [22] L. F. Sanchez-Peralta, J. B. Pagador, A. Picon, A. J. Calderon, F. Polo, N. Andracka, R. Bilbao, B. Glover, C. L. Saratxaga, and F. M. Sanchez-Margallo, “PIC-COLO White-Light and Narrow-Band Imaging Colonoscopic Dataset: A Performance Comparative of Models and Datasets,” *Applied Sciences*, vol. 10, no. 23, p. 8501, 2020.
- [23] A. Sanchez-Gonzalez, “Colonoscopy image pre-processing for the development of computer-aided diagnostic tools,” in *Colonoscopy and Colorectal Cancer Screening*, IntechOpen, 2017.
- [24] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.
- [25] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An image is worth 16x16 words: Transformers for image recognition at scale,” *arXiv preprint arXiv:2010.11929*, 2020.

- [26] A. Borji and L. Itti, “Bayesian optimization explains human active search,” in *Advances in Neural Information Processing Systems*, 2013. <https://proceedings.neurips.cc/paper/2013/file/a3f390d88e4c41f2747bfa2f1b5f87db-Paper.pdf>.
- [27] D. Chicco and G. Jurman, “The advantages of the matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation,” *BMC Genomics*, vol. 21, no. 1, p. 6, 2020.
- [28] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, “How transferable are features in deep neural networks?,” in *Advances in Neural Information Processing Systems*, vol. 27, 2014.
- [29] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” *arXiv preprint arXiv:1711.05101*, 2017.
- [30] PyTorch Contributors, “ReduceLROnPlateau,” 2026.
- [31] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A simple way to prevent neural networks from overfitting,” *Journal of Machine Learning Research*, vol. 15, no. 56, pp. 1929–1958, 2014.
- [32] L. Prechelt, “Early stopping – but when?,” in *Neural Networks: Tricks of the Trade*, vol. 1524 of *Lecture Notes in Computer Science*, pp. 55–69, Springer, 1998.
- [33] C. Shorten and T. M. Khoshgoftaar, “A survey on image data augmentation for deep learning,” *Journal of Big Data*, vol. 6, no. 60, 2019.
- [34] G. Jocher, A. Chaurasia, and J. Qiu, “Ultralytics YOLOv8,” 2023. AGPL-3.0 license.
- [35] Ultralytics, “YOLOv8,” 2023.
- [36] J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, and F. Vilariño, “WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians,” *Computerized Medical Imaging and Graphics*, vol. 43, pp. 99–111, 2015.
- [37] S. Sh, “YOLO_SAM2,” 2024.
- [38] Y. Zhang, P. Sun, Y. Jiang, D. Yu, F. Weng, Z. Yuan, P. Luo, W. Liu, and X. Wang, “ByteTrack: Multi-object tracking by associating every detection box,” in *Computer Vision – ECCV 2022*, 2022.
- [39] D. Malerba and V. Pasquadibisceglie, “Data-centric AI,” *Journal of Intelligent Information Systems*, vol. 62, no. 6, pp. 1493–1502, 2024.
- [40] Google DeepMind, “Modelos Gemini: una familia de modelos de IA generativa,” 2026.