Segmentación Semántica Binaria de Pólipos Colorrectales en Imágenes Endoscópicas mediante Arquitectura U-Net

Ivan Antonio Garcia Castillo
Alejandro Iñiguez Capitaine
Ingeniería Biomédica
Universidad Modelo
Mérida, Yucatán

***Resumen*—El cáncer colorrectal es una de las principales causas de muertes por neoplasia a nivel mundial. Aunque la colonoscopia es el estándar de oro para su detección, ésta presenta una tasa de omisión de pólipos de hasta un 22 % debido a factores humanos y ópticos. Ante esta problemática, durante los últimos años se han desarrollado sistemas computacionales basados en aprendizaje profundo, consiguiendo detectar, localizar e incluso segmentar con exactitud pólipos en una imagen. En este trabajo, se propone un modelo de segmentación semántica binaria de pólipos utilizando el dataset Kvasir-SEG, el cual consta de 1000 imágenes de colonoscopía. Se aplicaron algunas técnicas de preprocesamiento y aumento de datos y se entrenó una red U-Net, con una etapa codificadora ResNet34 preentrenada con ImageNet y una decodificadora UnetDecoder estándar, implementada en segmentation_models_pytorch. Al finalizar el entrenamiento se realizaron mediciones del conjunto de prueba incluyendo exactitud, coeficiente Dice promedio, puntaje F2 e índice de superposición, los cuales dieron como resultado las métricas 0.9700, 0.9039, 0.9068, y 0.8427, respectivamente, demostrando un rendimiento muy competitivo frente a otras investigaciones y estableciéndose como un instrumento eficaz para brindar asistencia clínica.**

***Abstract*—Colorectal cancer is one of the leading causes of neoplasm deaths worldwide. Although colonoscopy is the gold standard for its detection, it presents a polyp miss rate of up to 22% due to human and optical factors. In response to this problem, computer systems based on deep learning have been developed in recent years, managing to detect, localize, and even accurately segment polyps in an image. In this work, a binary semantic segmentation model for polyps is proposed using the Kvasir-SEG dataset, which consists of 1000 colonoscopy images. Some preprocessing and data augmentation techniques were applied, and a U-Net network was trained with a ResNet34 encoder stage pre-trained on ImageNet and a standard UnetDecoder, implemented in segmentation_models_pytorch. At the end of the training, measurements on the test set were performed, including accuracy, average Dice coefficient, F2 score, and intersection over union (IoU), which yielded the metrics 0.9700, 0.9039, 0.9068, and 0.8427, respectively, demonstrating a highly competitive performance compared to other research and establishing it as an effective instrument to provide clinical assistance.**

***Palabras clave*—Cáncer Colorrectal, Segmentación Semántica Binaria, Deep Learning, Pólipos, Colonoscopia, Transfer Learning**

I. Introducción y planteamiento del problema

El cáncer colorrectal (CCR) es una proliferación irregular de células en las zonas del colon o recto, que comienza como grupos pequeños de células llamados pólipos que suelen ser benignos, pero algunos pueden convertirse en tipos de cáncer colorrectal con el tiempo. Actualmente, es la tercera neoplasia más común y la segunda causa más común de muertes por cáncer a nivel mundial. Los informes epidemiológicos de 2022 mostraron que hubo 1,142,286 nuevos casos y 538,167 muertes en todo el mundo atribuibles a este tipo de cáncer [1]. Además, en México, la enfermedad también es una preocupación importante, con más de 16,000 casos y alrededor de 8,200 muertes vinculadas a este cáncer diagnosticadas durante 2022, y es la principal causa de muertes relacionadas con el cáncer en el país, superando incluso al cáncer de mama y de próstata [2].

La colonoscopia sigue siendo el método estándar de oro para el diagnóstico y la extirpación de pólipos. Sin embargo, a pesar de ser el procedimiento más robusto, es altamente dependiente de la habilidad del operador y del equipo óptico, pues éste puede presentar limitaciones técnicas significativas, tales como empañamiento, desenfoque y reflexión especular [3]. Estas dificultades visuales contribuyen a errores en el diagnóstico; de hecho, se ha reportado que la tasa de omisión de pólipos en la colonoscopia puede alcanzar hasta un 22 % [4], especialmente cuando las lesiones son planas o de un tamaño muy reducido.

Para superar estos desafíos, los sistemas de diagnóstico asistido por computadora (CAD) entrenados mediante deep learning se han posicionado como una herramienta fundamental para mejorar la tasa de detección y reducir la subjetividad clínica del operador [13]. No obstante, la principal limitación técnica de estos modelos sigue siendo su escasa capacidad de generalización. Al sobreajustarse a las propiedades ópticas de un único centro hospitalario o a un modelo específico de endoscopio, las redes experimentan una caída drástica en su rendimiento al enfrentarse a datos provenientes de entornos clínicos nuevos, lo que se conoce como cambio de dominio

(domain shift).

II. Objetivos

II-A. Objetivo general

Desarrollar e implementar un modelo de segmentación semántica basado en una arquitectura U-net con un codificador preentrenado, orientado a la delimitación precisa de pólipos colorrectales en imágenes endocópicas con el propósito de mitigar la pérdida de generalización y proporcionar una herramienta de asistencia al diagnóstico clínico.

II-B. Objetivos específicos

- **D**esarrollar un modelo de segmentación semántica basado en la arquitectura U-Net para mejorar la extracción de características multiescala.
- **D**iseñar e implementar un esquema de data augmentation apropiado para imágenes endoscópicas.
- **E**ntrenar el modelo U-Net+ResNet34 con función de pérdida combinada (Dice + BCE) y estrategia de regularización.
- **E**valuar el modelo con las métricas Dice, IoU, Precision, Recall y F2-score sobre el conjunto de prueba.

III. Revisión de la literatura

III-A. Métodos clásicos (pre-deep learning)

Actualmente, la aplicación del aprendizaje profundo para la detección, clasificación y segmentación de pólipos colorrectales, y también de imágenes médicas en general, ya está en práctica (es decir, el estándar). Sin embargo, anteriormente estos modelos utilizaban herramientas clásicas de procesamiento de imágenes como algoritmos de umbralización tradicionales y segmentación basada en textura, y agrupamiento Fuzzy C-Means (FCM) como métodos avanzados de procesamiento de imágenes, y algunos algoritmos de segmentación de pólipos. Estos enfoques detectaban áreas aberrantes por intensidad, color o características texturizadas en imágenes endoscópicas. Sin embargo, el rendimiento de estos métodos era débil, aunque estas técnicas fueron un gran avance en el área de procesamiento digital de imágenes, aunque fácilmente podíamos ver en una etapa temprana de procesamiento que el pólipo es altamente visualmente irregular, de color claro, reflejos de vidrio especular en el contexto de una colonoscopia, y las diferencias morfológicas en la presentación de lesiones dificultan que el rendimiento de muchas de las técnicas sea satisfactorio.

El Fuzzy C-Means (FCM) es un algoritmo de agrupamiento que da a cada píxel una probabilidad de pertenencia para segmentar. A diferencia de los métodos binarios, este algoritmo permite manejar la incertidumbre de los bordes de los pólipos y la transición suave entre tejido sano y anormal. Sin embargo, estos cálculos también dependían de pocas características de bajo nivel y eran extremadamente sensibles al ruido, variaciones de intensidad de luz y selección de parámetros iniciales, lo cual es muy importante para segmentar imágenes de colonoscopia. En Jha et al., FCM se empleó como técnica de referencia para la segmentación de pólipos, obteniendo métricas considerablemente bajas dentro de las arquitecturas modernas de aprendizaje profundo, con un DSC de 0.4997 y un IoU de 0.4356 [6].

Los métodos clásicos de umbralización también dividieron la región del pólipo basándose en diferencias de intensidad o color de las regiones de la imagen para determinar el fondo y el pólipo. Debido a los mismos problemas que los FCM, estos enfoques han mostrado una tasa de éxito muy baja en imágenes de colonoscopia reales, ya que al ser entornos predominantemente cubiertos de mucosa, el contraste entre el pólipo y el tejido sano y las variaciones de iluminación hicieron que su segmentación también fuera poco fiable. En contraste, se aplicaron técnicas basadas en textura para identificar áreas sospechosas basadas en estructuras espaciales y distribuciones estadísticas de píxeles. Sin embargo, las desventajas que mostraron fueron evidentes con respecto a pólipos pequeños, lesiones planas o estructuras anatómicas complejas [13]. En consecuencia, estas limitaciones impulsaron el cambio de pasar de pequeños sistemas de segmentación asistida por computadora a modelos de aprendizaje profundo que pueden extraer mejores características semánticas, mientras aprenden de datos médicos más masivos.

III-B. Redes Neuronales Convolucionales

Las Redes Neuronales Convolucionales (CNN) ofrecieron un avance significativo y temprano en el análisis de imágenes médicas, donde los métodos convencionales dependen de características extraídas manual y subjetivamente, mientras que las CNN permitieron el modelado automático de representaciones jerárquicas complejas a partir de píxeles de imagen. Basado en el uso de un diseño simétrico en forma de Üz la elección estratégica de conexiones de salto, una innovación arquitectónica que ha ganado popularidad en la etapa reciente y que también se ha aplicado para propósitos de segmentación ha sido la arquitectura U-net [14]. Estas conexiones son cruciales porque al hacer posible integrar el conocimiento semántico profundo del codificador, con los mapas de características de alta resolución espacial proporcionados por el decodificador necesarios para reconstruir los bordes difusos de estructuras como los pólipos. A pesar del éxito del U-net tradicional, investigaciones recientes encontraron que también es posible optimizar la capacidad de extracción de U-net. Para proporcionar representaciones más grandes y prevenir el fenómeno de desaparición del gradiente en modelos profundos, varios autores han reemplazado el codificador original con redes residuales preentrenadas (como la familia ResNet [8], [10]) para obtener modelos híbridos más robustos para variaciones clínicas en características visuales.

III-C. Arquitecturas Basadas en Transformers

Por otro lado, los avances recientes han explorado el uso de arquitecturas basadas en Transformers, como lo son los casos de TransFuse [12] y Polyp-PVT [18]. La idea principal de estos modelos es combinar la capacidad de las CNN para extraer detalles locales con la habilidad de los Transformers para

entender todo el contexto global de la imagen endoscópica. Aunque estas arquitecturas logran métricas sumamente altas (con coeficientes Dice superiores a 0.92 en el dataset Kvasir-SEG), presentan un obstáculo clínico importante: requieren costos computacionales significativamente mayores y bases de datos mucho más masivas para poder entrenarse de manera efectiva en comparación con los métodos tradicionales basados puramente en convoluciones.

III-D. Arquitecturas Codificador-Decodificador para la Segmentación de Pólipos

Ronneberger et al. propusieron una arquitectura U-Net [15] que estableció la arquitectura encoder-decoder con conexiones de salto como el marco predominante para la segmentación de imágenes médicas. En el contexto de la segmentación de pólipos, Jha et al. [7] presentaron el conjunto de datos Kvasir-SEG y evaluaciones de referencia basadas en los estándares U-Net y ResUNet, y también resultaron con coeficientes Dice de 0.818 y 0.714, respectivamente, en el benchmark de 1,000 imágenes. Posteriormente, el grupo propuso ResUNet++ [17], integrando bloques residuales, módulos de compresión y excitación, mecanismos de atención y agrupamiento piramidal espacial atrous, demostrando un rendimiento significativamente mejor en comparación con la línea base ResUNet. Zhou et al. [16] propusieron UNet++, que rediseña las conexiones de salto como subredes densas anidadas que reducen progresivamente la brecha semántica entre los mapas de características del encoder y el decoder, y reportaron aumentos consistentes sobre el U-Net original en múltiples tareas de segmentación médica.

III-E. Enfoques de Atención y Multiescala

Para mejorar la detección de lesiones pequeñas o difíciles, otros enfoques han incorporado mecanismos de atención. Por ejemplo, la arquitectura PraNet [11] emplea un sistema de atención inversa que le enseña al modelo a refinar los bordes de la predicción y descartar el ruido del tejido sano, alcanzando un Dice de 0.8982. En una línea similar, pero enfocada en la transparencia médica, Asare et al. [19] le agregaron interpretabilidad a una U-Net+ResNet34 utilizando mapas de calor (GradCAM). De esta forma, además de lograr métricas competitivas (Dice de 0.8902), el modelo le ofrece al médico una justificación visual de qué características anatómicas fueron las que determinaron la segmentación del pólipo.

III-F. Transfer Learning y Backbones Ligeros

Como se muestra a lo largo de la literatura, el preentrenamiento de imágenes para la inicialización del encoder en ImageNet ha demostrado una efectividad significativa en la segmentación de imágenes médicas. Se encontró que un nuevo U-Net con encoder MobileNetV2 alcanza 0.8971 Dice en Kvasir-SEG [36] mientras logra un rendimiento similar con variantes más pesadas basadas en ResNet50 propuestas por Jha et al. (Dice $= 0,8154$, IoU $= 0,7396$), confirmando que el preentrenamiento del encoder es una medida importante, más que la capacidad bruta. Un artículo reciente en Zenodo mostró que un U-Net estándar entrenado durante 100 épocas logró un Dice de 0.9449 en Kvasir-SEG, indicando que conjuntos de entrenamiento más grandes y estrategias de aumento pueden cerrar la brecha entre arquitecturas con arquitecturas complejas y arquitecturas sin alterar.

Finalmente, un factor que ha demostrado ser determinante en la literatura es el Transfer Learning mediante la inicialización de codificadores con ImageNet. Diversos estudios han confirmado que, ya sea utilizando backbones ligeros como MobileNetV2 [20] o arquitecturas más profundas como ResNet50 [17], arrancar con pesos preentrenados evita el sobreajuste y acelera la convergencia de manera más efectiva que simplemente aumentar la capacidad bruta de la red. De hecho, investigaciones recientes [37] sugieren que utilizar estrategias sólidas de aumentación de datos junto con un buen preentrenamiento puede hacer que una U-Net estándar logre métricas a la par de arquitecturas mucho más pesadas o complejas. Para superar esta limitación, el objetivo de este trabajo es introducir un algoritmo de segmentación semántica binaria propuesta para el dataset Kvasir-SEG [5] para resolver esta tarea.

IV. Metodología detallada del procesamiento de imágenes

IV-A. Descripción del dataset

En este trabajo se utilizó el dataset Kvasir-SEG, propuesto por Jha et al. [7], el cual es conocido por ser uno de los benchmarks más utilizados para segmentación semántica de pólipos colorrectales en imágenes de colonoscopía. Este dataset contiene 1000 imágenes RGB obtenidas durante procedimientos reales de colonoscopía, acompañadas de sus respectivas máscaras binarias anotadas manualmente por especialistas clínicos. Las imágenes presentan una alta variabilidad en características como tamaño del pólipo, iluminación, contraste, textura y presencia de artefactos endoscópicos, lo que representa un escenario variado y realista para modelos de segmentación automática. En la 1 se pueden observar algunas de estas imágenes con sus respectivas máscaras.

Previo al entrenamiento, se pudo notar que algunos frames de colonoscopía del dataset contenían overlays clínicos muy notorios, mapas de navegación endoscópica a color u otros artefactos que pudieran introducir sesgos durante el aprendizaje. De manera similar, se observaron casos en los que ciertos componentes anatómicos del pólipo estaban ocultos por otras estructuras, como recortes en un mapa de navegación. Tales escenarios podrían sesgar el modelo durante el entrenamiento y, lo que es más importante, llevar a una mala generalización, especialmente considerando artefactos visuales y regiones parcialmente ocultas en las imágenes endoscópicas [35].

Para caracterizar las estadísticas de resolución y la distribución de la clase de primer plano antes del desarrollo del modelo, se realizó un análisis de datos en una muestra aleatoria de $n = 200$ pares de imágenes. La Fig. 2 resume los principales hallazgos encontrados. El diagrama de dispersión de resolución (panel izquierdo) revela que la gran mayoría de las imágenes se agrupa en el rango de 480–620 píxeles en la dimensión vertical y 500–650 píxeles en

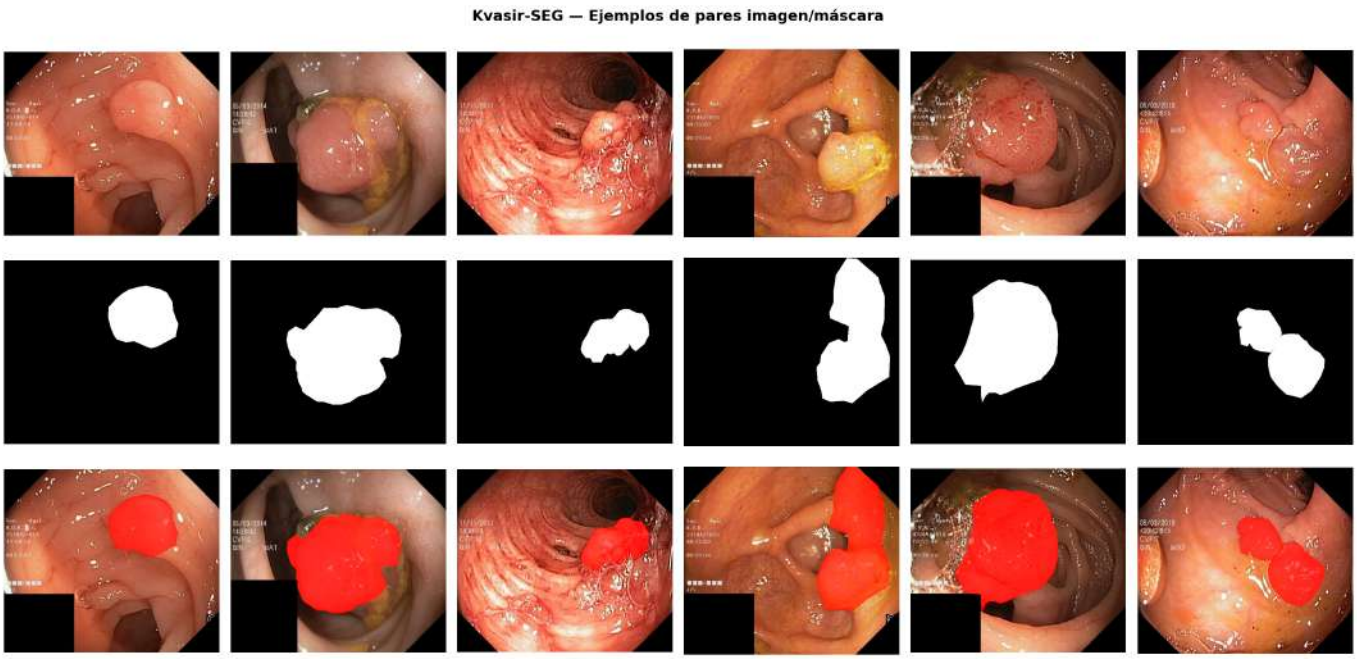

Figura 1. Pares de imagen-máscara del dataset Kvasir-SEG. Fila superior: frames de endoscopía originales. Fila central: Máscaras de verdad fundamental (ground truth) anotadas por expertos. Fila inferior: Superposición de máscaras en rojo sobre su respectivo frame original. Estas muestras revelan variación de tamaño, iluminación y textura de pólipos, entre otras características.

la dimensión horizontal, con solo unos pocos valores atípicos de alta resolución extremos que alcanzan hasta $1,067 \times 1,495$ píxeles. indica la dimensión de redimensionamiento utilizada para el entrenamiento como objetivo, y confirma que todas las imágenes pierden información de textura, pero genera una cantidad considerable de ahorro computacional.

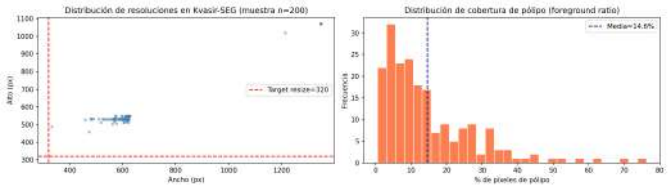

Figura 2. Análisis exploratorio de datos del conjunto de datos Kvasir-SEG ($n = 200$). Izquierda: diagrama de dispersión de la resolución de imagen (ancho vs. alto en píxeles), con las líneas rojas discontinuas marcando la dimensión de redimensionamiento objetivo de 320×320 píxeles. Derecha: histograma del porcentaje de píxeles de pólipo por imagen, con la línea azul discontinua indicando la media del conjunto de datos del 14,6 %.

El histograma de cobertura del primer plano (panel derecho) revela una distribución sesgada a la derecha con una razón media de primer plano del 14,6 %. La mayoría de las imágenes contienen pólipos que ocupan menos del 20 % del área total de píxeles, con una cola larga que se extiende hasta aproximadamente el 75 %. Este desequilibrio de clases, aunque moderado en comparación con otras tareas de segmentación médica, motiva el uso del Dice Loss como función objetivo principal, dada su invarianza ante la frecuencia absoluta de clases primer plano-fondo [31]. La amplia varianza en el tamaño del pólipo justifica además la agregación de características multiescala inherente al diseño de skip connections de U-Net.

El dataset fue dividido en tres splits: entrenamiento, validación y prueba, utilizando una proporción de 80/10/10, correspondiente a 800, 100 y 100 imágenes con sus respectivas máscaras, estableciendo una semilla aleatoria fija de 42 para garantizar la reproducibilidad. Esta estrategia es ampliamente utilizada en entrenamiento de redes neuronales para procesamiento de imágenes médicas, debido a que permite maximizar la cantidad de muestras disponibles para el entrenamiento sin comprometer la calidad de evaluación del modelo [7]. Adicionalmente, el split de validación fue empleado para monitorear la convergencia del modelo y aplicar técnicas de regularización como early stopping, mientras que el split de prueba permaneció completamente aislado durante el entrenamiento para evitar fuga de información (data leakage).

IV-B. Arquitectura: U-Net con Codificador ResNet34

El modelo propuesto sigue el paradigma U-Net codificador-decodificador [15] con un backbone ResNet34 [9] sustituyendo la trayectoria de contracción original. ResNet34 comprende un bloque convolucional inicial seguido de cuatro etapas residuales que producen mapas de características a $1/2$, $1/4$, $1/8$, $1/16$ y $1/32$ de la resolución espacial de entrada. Cada bloque residual implementa la formulación de skip connection:

$$\mathbf{h}_l = \mathcal{F}(\mathbf{h}_{l-1}; \mathbf{W}_l) + \mathbf{h}_{l-1} \tag{1}$$

donde $\mathcal{F}(\cdot; \mathbf{W}_l)$ denota el mapeo residual compuesto por dos bloques de convolución-BN-ReLU de 3×3, y el atajo

aditivo evita la degradación del gradiente en profundidad. El codificador se inicializa con pesos preentrenados en ImageNet [33], lo que permite la transferencia de representaciones de bordes, color y textura de bajo nivel aprendidas a partir de 1,2 millones de imágenes naturales al análisis de colonoscopia.

El decodificador utiliza cuatro pasos de muestreo ascendente, con un paso de muestreo ascendente bilineal de $2\times$ seguido de un bloque convolucional para la concatenación del mapa de características, muestreado ascendentemente con el salto del codificador. Este proceso retiene el detalle espacial fino que de otro modo se perdería al reducir la muestra, y ha sido especialmente exitoso en delinear bordes irregulares o difusos asociados con pólipos sésiles [16]. La salida final del decodificador pasa por una capa convolucional 1×1 seguida de una activación sigmoide para producir un mapa de probabilidad a nivel de píxel $\hat{\mathbf{p}} \in [0,1]^{H\times W}$.

IV-C. Función de Pérdida

El objetivo de entrenamiento combina la pérdida Binary Cross-Entropy (BCE) y el Dice Loss en igual proporción, una formulación que promueve simultáneamente la exactitud de clasificación por píxel y la superposición global de regiones:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(\hat{p}_i) + (1-y_i)\log(1-\hat{p}_i)\right] \tag{2}$$

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2\sum_{i=1}^{N} y_i \hat{p}_i + \varepsilon}{\sum_{i=1}^{N} y_i + \sum_{i=1}^{N} \hat{p}_i + \varepsilon} \tag{3}$$

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{BCE}} + \mathcal{L}_{\text{Dice}} \tag{4}$$

donde N es el número total de píxeles en el lote, $y_i \in \{0,1\}$ es la etiqueta de verdad fundamental, $\hat{p}_i \in [0,1]$ es la probabilidad predicha, y $\varepsilon = 10^{-6}$ es una constante de suavizado que evita la división por cero. Decidimos combinar ambas funciones de pérdida (Dice + BCE) porque se complementan perfectamente para este problema clínico. Por un lado, la BCE nos ayudó a mantener la estabilidad al evaluar las predicciones píxel por píxel durante las primeras etapas. Por otro lado, dado que en una colonoscopia los pólipos suelen ocupar una porción muy pequeña de la imagen frente a una gran cantidad de tejido sano (generando un fuerte desbalance de clases), incorporar la pérdida Dice obligó al modelo a priorizar la delimitación exacta de la lesión sin dejarse engañar por el exceso de fondo. [31].

IV-D. Métricas de Evaluación

El rendimiento del modelo se evaluó en el conjunto de prueba del dataset, utilizando las siguientes métricas derivadas de predicciones binarias a nivel de píxel obtenidas al aplicar un umbral a $\hat{\mathbf{p}}$ en $\tau = 0,5$.

El coeficiente Dice (equivalente a la puntuación F1) se define como:

$$\text{Dice} = \frac{2\,\text{TP}}{2\,\text{TP} + \text{FP} + \text{FN}} \tag{5}$$

La Intersección sobre Unión (IoU), también conocida como índice de Jaccard, es:

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \tag{6}$$

La Precisión y el Recall se definen como:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \qquad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{7}$$

La puntuación F2 pondera el Recall el doble que la Precisión, reflejando el mayor costo clínico de las detecciones perdidas (falsos negativos) en relación con las falsas alarmas:

$$F_2 = \frac{(1+4)\,\text{TP}}{(1+4)\,\text{TP} + 4\,\text{FN} + \text{FP}} \tag{8}$$

La Exactitud a nivel de píxel es:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{9}$$

IV-E. Preprocesamiento de Datos y Aumentación

Como primer paso del preprocesamiento, todas las imágenes y máscaras fueron redimensionadas a una resolución espacial de 320×320 píxeles. Esta resolución fue seleccionada debido a que representa un compromiso adecuado entre costo computacional y preservación de detalles anatómicos relevantes, además de ser consistente con resoluciones comúnmente utilizadas en benchmarks de segmentación de pólipos [7], [11]. Asimismo, la divisibilidad por $2^5 = 32$ facilita las operaciones sucesivas de downsampling y upsampling realizadas por arquitecturas encoder-decoder basadas en U-Net.

Para las imágenes RGB se utilizó interpolación bicúbica durante el redimensionamiento, ya que este método preserva de mejor manera las transiciones suaves de intensidad y reduce artefactos visuales en comparación con interpolaciones de menor orden. Por otro lado, las máscaras binarias fueron redimensionadas mediante interpolación nearest-neighbor, evitando la introducción de valores intermedios no válidos entre clases durante el proceso de reescalado.

Posteriormente, las imágenes fueron normalizadas utilizando la media y desviación estándar del dataset ImageNet, específicamente $\mu = [0{,}485, 0{,}456, 0{,}406]$ y $\sigma = [0{,}229, 0{,}224, 0{,}225]$. Esta normalización se empleó debido a que el encoder ResNet34 utilizado en el presente trabajo fue inicializado con pesos preentrenados sobre ImageNet [34]. Aunque las imágenes endoscópicas presentan una distribución cromática predominantemente rojiza y rosada, el uso de estadísticas de ImageNet permite mantener compatibilidad estadística con las activaciones esperadas por la red preentrenada, favoreciendo la estabilidad numérica y la transferencia efectiva de características visuales generales aprendidas durante el preentrenamiento.

Con el objetivo de mejorar la capacidad de generalización del modelo y reducir el sobreajuste asociado al tamaño limitado del dataset, se implementó una estrategia de data

augmentation utilizando la librería Albumentations [24]. Las transformaciones fueron aplicadas únicamente durante la fase de entrenamiento, mientras que los conjuntos de validación y prueba conservaron exclusivamente operaciones de preprocesamiento determinísticas.

Las transformaciones geométricas incluyeron volteo horizontal, volteo vertical, rotación aleatoria de $90°$, así como operaciones de traslación, escalamiento y rotación mediante ShiftScaleRotate. Estas transformaciones permiten modelar variaciones plausibles asociadas a cambios de orientación del endoscopio durante la exploración clínica. Adicionalmente, se utilizó ElasticTransform para simular deformaciones suaves de tejido biológico, incrementando la robustez del modelo frente a variaciones anatómicas presentes en procedimientos reales de colonoscopía.

En el dominio fotométrico se aplicaron transformaciones de brillo y contraste aleatorio, desenfoque gaussiano y adición de ruido gaussiano. Estas operaciones buscan modelar condiciones reales de adquisición, incluyendo variaciones de iluminación, desenfoque por movimiento y ruido inducido por el sistema óptico del endoscopio. Finalmente, se empleó CoarseDropout como mecanismo de regularización adicional, simulando oclusiones parciales y obligando al modelo a aprender representaciones espaciales más estables.

El uso de data augmentation en segmentación médica ha demostrado mejorar significativamente la generalización y reducir el sobreajuste en datasets pequeños [25], incluso si no está enfocado directamente en sumar más volumen de datos para la fase de entrenamiento. Entrando en detalles técnicos, la secuencia de aumentación utilizada, ilustrada en la Fig. 3, incluye: volteo horizontal (probabilidad del 50 %), volteo vertical (probabilidad del 50 %), rotación aleatoria en $[-30, +30]$, escalado aleatorio por un factor extraído uniformemente de $[0, 8, 1, 2]$, jitter de color que modifica el brillo $(\pm 0, 2)$, el contraste $(\pm 0, 2)$, la saturación $(\pm 0, 2)$ y el matiz $(\pm 0, 1)$, y desenfoque gaussiano con un kernel de 3×3 (probabilidad del 30 %). De igual forma, no está de más señalar que todas las transformaciones espaciales se aplican de forma idéntica a la imagen y a su máscara correspondiente para preservar la alineación de la ground truth.

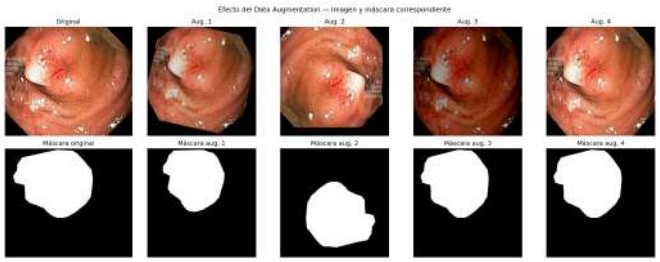

Figura 3. Ejemplos del efecto del pipeline de aumentación de datos sobre una muestra de entrenamiento representativa. Fila superior: versiones aumentadas del fotograma endoscópico original (columnas 2–5) junto al original (columna 1), ilustrando la variación geométrica y fotométrica introducida durante el entrenamiento. Fila inferior: las máscaras binarias transformadas de forma correspondiente, confirmando la aumentación conjunta coherente de imagen y máscara.

IV-F. Protocolo de Entrenamiento

El modelo se entrenó durante un máximo de 50 épocas con un tamaño de lote de 16, empleando el optimizador AdamW [22] con una tasa de aprendizaje inicial de $\eta_0 = 10^{-4}$ y una decaída de pesos de $\lambda = 10^{-2}$. AdamW desacopla la decaída de pesos de la actualización basada en gradientes, proporcionando una regularización más fundamentada que la formulación Adam con penalización L_2 y generalmente produciendo una mejor generalización en entornos de transfer learning [22].

La tasa de aprendizaje se atenuó siguiendo un calendario de coseno:

$$\eta_t = \eta_{\text{mín}} + \frac{1}{2}(\eta_0 - \eta_{\text{mín}}) \left(1 + \cos\left(\frac{t\,\pi}{T}\right)\right) \tag{10}$$

donde t es la época actual, $T = 50$ es el periodo, y $\eta_{\text{mín}} = 10^{-6}$ es la tasa de aprendizaje mínima. El cosine annealing reduce la tasa de aprendizaje de forma suave desde η_0 hasta $\eta_{\text{mín}}$, facilitando una convergencia estable en las etapas finales del entrenamiento y evitando las mesetas abruptas de pérdida asociadas con los calendarios de decaída escalonada [23].

Adicionalmente, implementamos un gradient clipping (recorte de gradientes) con un valor máximo de 1.0. Esta medida de seguridad fue necesaria porque, al tener el codificador preentrenado y un decodificador iniciando desde cero, queríamos evitar que los gradientes se dispararan y desestabilizaran los pesos de la red durante las primeras iteraciones.

Todo el pipeline fue implementado utilizando PyTorch. El modelo fue entrenado en una única GPU-T4, en un entorno de ejecución de Google Colab, y el encoder ResNet34 se inicializó con pesos preentrenados en ImageNet mediante la librería segmentation-models-pytorch.

V. Resultados

V-A. Dinámica del Entrenamiento

Las curvas de pérdida y métricas de entrenamiento y validación a lo largo de las 33 épocas efectivas se muestran en la

Fig. 4. La pérdida combinada (panel izquierdo) muestra una disminución monotónica pronunciada desde aproximadamente 0,54 (época 1) hasta aproximadamente 0,09–0,10 (época final) tanto para las curvas de entrenamiento como las de validación, con la pérdida de validación consistentemente por debajo del training loss a partir de la época 5. Este ordenamiento refleja el comportamiento estándar de los modelos que emplean dropout y aumentación: el paso forward de validación utiliza inferencia determinista sin regularización estocástica, y las estadísticas de normalización en tiempo de prueba de las capas BatchNorm (que incorporan promedios acumulados del entrenamiento) pueden producir una pérdida aparente menor en el conjunto de validación durante las fases iniciales [32]. La divergencia de training loss-validación no es persistente con respecto a los datos de entrenamiento, lo que sugiere que no ocurrió sobreajuste estructural durante la ejecución del entrenamiento.

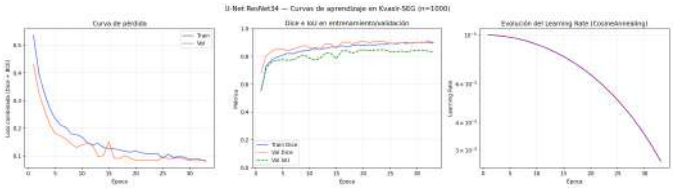

Figura 4. Dinámica de entrenamiento del modelo U-Net ResNet34 en Kvasir-SEG ($n = 1,000$). Izquierda: pérdida combinada (BCE + Dice) en función de la época para las particiones de entrenamiento (azul) y validación (naranja). Centro: evolución del coeficiente Dice para entrenamiento (azul) y validación (naranja), e IoU de validación (verde discontinuo). Derecha: calendario de tasa de aprendizaje de cosine annealing desde $\eta_0 = 10^{-4}$ hasta $\eta_{\min} = 10^{-6}$ a lo largo de 50 épocas.

Las curvas de Dice e IoU muestran una rápida mejora en las métricas en las primeras 5-7 épocas, antes de haber alcanzado una fase estable de aproximadamente 0.90 en el Dice de entrenamiento, mientras que el Dice de validación muestra valores de convergencia similares en las primeras 5-7 épocas. El cambio abrupto temporal en el Dice de validación alrededor de las épocas 13-15 con una perturbación en la curva de pérdida también sugiere un período inestable en el medio del programa de tasa de cosine annealing, cuando η_t es demasiado grande para algunas actualizaciones de peso y cuando el rendimiento de validación se ve comprometido después de un corto período de tiempo. El programa de cosine annealing (panel derecho) confirma la decaída suave de la tasa de aprendizaje desde 10^{-4} hasta aproximadamente $3,0 \times 10^{-5}$ en la época final, correspondiente a la sección de la curva de coseno donde la tasa de decaída es mayor.

V-B. Resultados Cuantitativos

La Tabla I reporta el conjunto completo de métricas de evaluación calculadas sobre la partición de prueba de 100 muestras. El modelo alcanza un coeficiente Dice medio de 0,9039 (±0,1239), un IoU de 0,8427 (±0,1619), una precisión de 0,9221 (±0,1055), un recall de 0,9130 (±0,1565), una puntuación F2 de 0,9068 (±0,1410) y una exactitud a nivel de píxel de 0,9700 (±0,0528).

Cuadro I
RENDIMIENTO EN EL CONJUNTO DE PRUEBA DEL MODELO U-NET RESNET34 PROPUESTO EN KVASIR-SEG ($n_{\text{TEST}} = 100$)

Métrica	**Media**	**Desv. Est.**	**Mín.**	**Máx.**
Dice	0,9039	0,1239	0,2329	0,9846
IoU	0,8427	0,1619	0,1318	0,9696
Precision	0,9221	0,1055	0,4732	1,0000
Recall	0,9130	0,1565	0,1410	1,0000
F2-score	0,9068	0,1410	0,1674	0,9880
Accuracy	0,9700	0,0528	0,7247	0,9979

La Fig. 5 grafica la distribución de las puntuaciones Dice por muestra y la distribución entre métricas. El histograma de Dice (panel izquierdo) presenta una distribución fuertemente sesgada a la izquierda con mediana 0,9565, lo que indica que el modelo se desempeña muy bien en la mayoría de las muestras de prueba, mientras que un pequeño grupo de casos difíciles desplaza la media hacia abajo. Específicamente, el histograma muestra una marcada concentración de puntuaciones en el intervalo $[0,90, 1,00)$ (aproximadamente el 80 % de las muestras), con una cola aproximándose hasta el mínimo de 0,2329. La brecha entre la media (0,9039) y la mediana (0,9565) señala que la media global presenta poca dificultad en los casos.

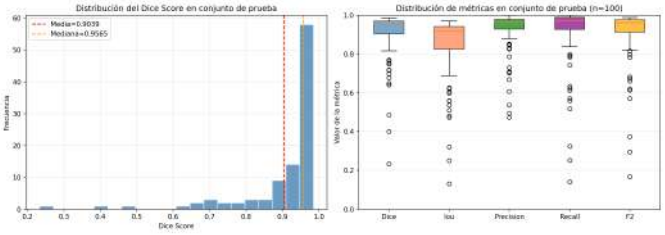

Figura 5. Distribución de las métricas de segmentación en el conjunto de prueba ($n = 100$). Izquierda: histograma de las puntuaciones Dice por muestra con la media (línea roja discontinua, 0,9039) y la mediana (línea naranja discontinua, 0,9565) indicadas. Derecha: diagramas de caja para Dice, IoU, Precisión, Recall y puntuación F2, ilustrando los rangos intercuartílicos, la estructura de valores atípicos y la variabilidad específica por métrica.

Los diagramas de caja (panel derecho) ofrecen información complementaria sobre la estructura distribucional por métrica. El IoU presenta el rango intercuartílico más amplio, lo que es consistente con su relación no lineal con el Dice (IoU = Dice / (2 − Dice)), que amplifica pequeñas diferencias de Dice en diferencias de IoU mayores a niveles de rendimiento más bajos. La Precisión muestra el rango intercuartílico más estrecho de las cinco métricas, lo que refleja la tendencia general del modelo hacia predicciones conservadoras (alta especificidad) en lugar de liberales (alta sensibilidad). Las cajas de Recall y puntuación F2 son más amplias, indicando que las tasas de falsos negativos son más variables entre las muestras de prueba que las tasas de falsos positivos, un patrón consistente con los errores de infrasegmentación observados cualitativamente en la Sección V-D.

V-C. Comparación con el Estado del Arte

La Tabla II presenta una comparación cuantitativa del modelo propuesto frente a métodos representativos evaluados en Kvasir-SEG. Los métodos se ordenan por coeficiente Dice ascendente.

Cuadro II
COMPARACIÓN CON MÉTODOS DEL ESTADO DEL ARTE EN KVASIR-SEG

Método	Dice	IoU	Recall	Precision
FCM (agrupamiento) [7]	0,4997	0,4356	—	—
ResUNet [7]	0,7139	0,6226	—	—
U-Net+ResNet50 [17]	0,8154	0,7396	0,8533	0,8532
U-Net+ResNet34+GradCAM [19]	0,8902	0,8023	0,9058	0,9083
U-Net+MobileNetV2 [36]	0,8971	0,8164	—	—
El nuestro (U-Net+ResNet34)	**0,9039**	**0,8427**	**0,9130**	**0,9221**
U-Net estándar 100 ép [37]	0,9449	0,9084	0,9584	0,9404

El modelo propuesto supera a la línea base U-Net+ResNet50 de Jha et al. [17] por un margen sustancial de 8,8 puntos de Dice y 10,3 puntos de IoU, confirmando el beneficio del transfer learning desde ImageNet frente al entrenamiento desde cero con un codificador más pesado. Supera a la U-Net ResNet34 con GradCAM de Asare et al. [19] (+1,4 puntos de Dice, +4,0 puntos de IoU) y a la variante respaldada por MobileNetV2 [36] (+0,7 puntos de Dice, +2,6 puntos de IoU). La U-Net estándar entrenada durante 100 épocas [37] reporta valores más altos (Dice 0,9449, IoU 0,9084), pero esta comparación debe interpretarse con cautela dado las posibles diferencias en la definición de la partición de entrenamiento-prueba, los regímenes de aumentación y las convenciones de reporte entre repositorios.

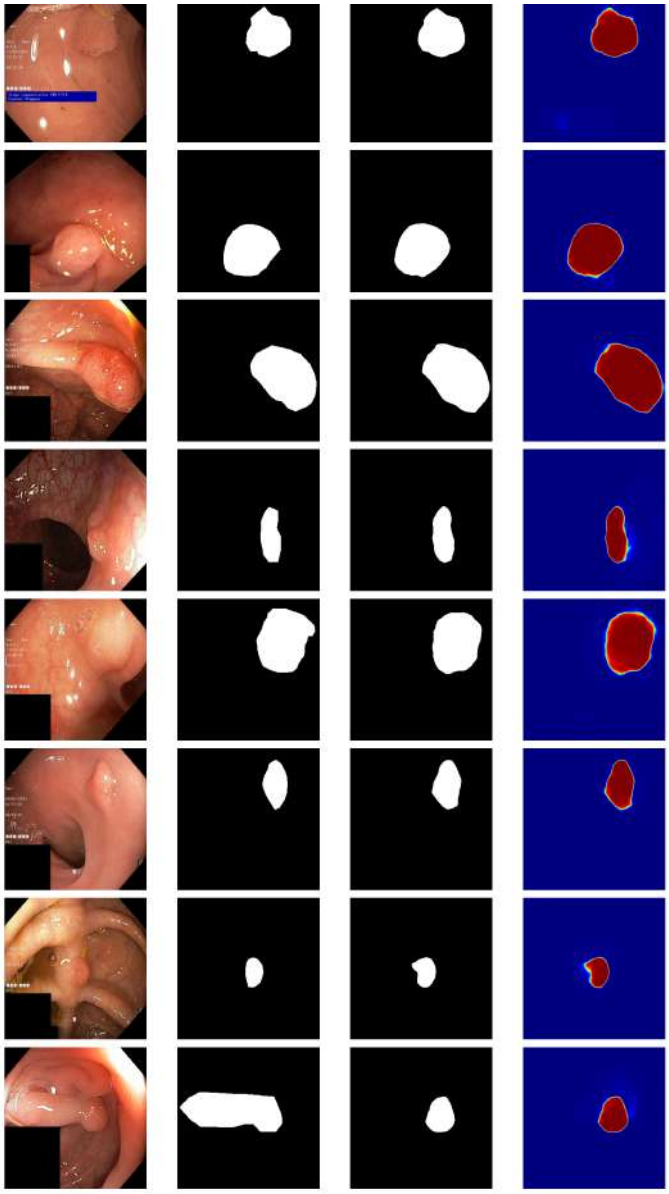

Figura 6. Resultados cualitativos de segmentación en el conjunto de prueba, ordenados por puntuación Dice por muestra. Columnas de izquierda a derecha: imagen endoscópica original, máscara binaria de verdad fundamental, máscara binaria predicha (umbral $\tau = 0,5$) y mapa de probabilidad sigmoide. Las filas 1-4 corresponden a los cuatro casos con mayor puntuación; las filas 5-8, a los cuatro con menor puntuación.

V-D. Análisis Cualitativo

La Fig. 6 presenta un análisis cualitativo de las predicciones del modelo para los cuatro casos de mejor rendimiento (filas 1-4) y los cuatro de peor rendimiento (filas 5-8) del conjunto de prueba, clasificados por coeficiente Dice por muestra. Cada fila muestra el fotograma endoscópico original, la máscara de verdad fundamental anotada por el experto, la máscara predicha por el modelo y el mapa de probabilidad sigmoide correspondiente.

En los casos de alto rendimiento (filas 1-4), las máscaras predichas replican los contornos de los ground truth respectivos, segmentando correctamente los bordes. Los mapas de probabilidad asignan de forma consistente valores de alta confianza (rojo intenso en la escala de pseudocolor) al interior de la región del pólipo, con una región de probabilidad intermedia en el borde. Esto sugiere que el modelo ha aprendido con éxito una representación del tejido poliposo que es robusta ante reflexiones especulares y gradientes sutiles de color en el borde de la lesión.

Los casos de fallo (filas 5-8) revelan varios modos de error característicos. La fila 5 muestra un caso donde la máscara

de ground truth cubre una región inusualmente alargada, y la máscara predicha captura solo el lóbulo central, segmentando incorrectamente el pólipo completo y fallando al encontrar bordes. Este patrón de error sugiere una capacidad limitada del modelo para manejar morfologías de pólipos alargadas o irregulares que se desvían de las formas convexas más usuales. La fila 8 demuestra un caso de fallo de segmentación casi completo (Dice mínimo de 0,2329), donde el ground truth contiene una lesión irregular grande con bajo contraste respecto a la mucosa. La máscara predicha captura solo una pequeña región central, probablemente porque las características del borde en esta muestra caen por debajo del umbral de decisión. Los mapas de probabilidad en los casos de fallo muestran gradientes de confianza menos pronunciados y menor concentración espacial, indicando incertidumbre genuina del modelo en lugar de una clasificación errónea en un límite de decisión difícil.

VI. Discusión y análisis crítico de resultados

VI-A. Análisis de la Generalización y la Estabilidad del Entrenamiento

En la Fig. 4 se puede observar el comportamiento de convergencia de aprendizaje, que indica que la combinación de transfer learning, cosine annealing y aumentación, produce un régimen de entrenamiento estable y bien regularizado. La ausencia de una brecha persistente entrenamiento-validación en las curvas de pérdida a lo largo de 33 épocas, a pesar del tamaño limitado del conjunto de datos de 800 muestras de entrenamiento, sugiere que el pipeline de aumentación diversifica suficientemente la distribución de entrenamiento para evitar que el codificador colapse en representaciones sobreajustadas al conjunto de datos específico. El calendario de cosine annealing contribuye a esta estabilidad al evitar las mesetas abruptas de pérdida asociadas con las políticas de decaída escalonada agresiva, permitiendo al optimizador continuar progresando en el régimen de parámetros de grano fino durante las últimas épocas de entrenamiento.

La inestabilidad temporal alrededor de las épocas 13–15, visible tanto en las curvas de pérdida de validación como en las de Dice de validación, es atribuible a la interacción entre la varianza de gradiente inducida por la aumentación y la tasa de aprendizaje moderadamente grande en el segmento medio del calendario de coseno. Aunque esta perturbación no persiste y no impide que el modelo se recupere hasta su mejor punto de control, motiva considerar estrategias de calentamiento (warm-up) o calendarios de período más largo en trabajos futuros.

VI-B. Interpretación de la Distribución de Métricas

La brecha entre el Dice medio (0,9039) y el Dice mediano (0,9565) tiene importantes implicaciones para el despliegue clínico. Una amplia mayoría de los casos se gestiona con una exactitud casi perfecta, pero un pequeño subconjunto de casos estructuralmente desafiantes —predominantemente lesiones grandes, difusas, multilobulares o de bajo contraste— es responsable de la desviación de la media respecto a la mediana. En un contexto de asistencia clínica, esta distribución argumenta a favor de incluir una señal de incertidumbre derivada del mapa de probabilidad sigmoide para señalar los casos donde la confianza del modelo es baja y se requiere revisión manual. Los mapas de probabilidad bien calibrados observados en la Fig. 6 apoyan este enfoque: la coherencia espacial y los gradientes de confianza de los mapas de probabilidad siguen cualitativamente el rendimiento Dice del caso correspondiente.

VI-C. Impacto de la Aumentación de Datos

La estrategia de aumentación de datos descrita en la Sección IV.D parece haber sido eficaz para prevenir el sobreajuste, dada la ausencia de una brecha divergente entrenamiento-validación. El pipeline de aumentación introduce variantes de entrenamiento geométrica y fotométricamente diversas de cada muestra, multiplicando efectivamente el tamaño aparente del conjunto de datos y reduciendo el riesgo de que el modelo memorice texturas de fondo o orientaciones específicas. Sin embargo, el pipeline de aumentación no incluye deformación elástica ni eliminación brusca (coarse dropout), estrategias que han demostrado mejorar aún más la generalización en el entrenamiento de U-Net para imágenes médicas [15]. La ausencia de estas transformaciones puede explicar en parte la vulnerabilidad residual ante pólipos morfológicamente atípicos.

VI-D. Análisis de Modos de Fallo e Infrasegmentación

El modo de fallo predominante observado en el análisis cualitativo es la infrasegmentación: el modelo tiende a predecir un subconjunto de la región de verdad fundamental en lugar de extender predicciones falsas más allá del borde verdadero. Esto es consistente con la Precisión ligeramente mayor (0,9221) que el Recall (0,9130) observados en la Tabla I, lo que indica que el modelo se inclina conservadoramente hacia los falsos negativos en lugar de los falsos positivos. Desde el punto de vista clínico, la infrasegmentación es el tipo de error más preocupante, ya que implica que partes del tejido lesionado no se resaltan para la atención del endoscopista. La métrica de puntuación F2 (0,9068), que penaliza los falsos negativos más severamente que los falsos positivos, proporciona una medida que refleja mejor esta asimetría en el costo clínico.

El análisis morfológico de los casos de peor rendimiento sugiere que los principales factores de infrasegmentación son: (i) regiones de pólipo con color y textura similares a la mucosa circundante; (ii) lesiones grandes que se extienden hasta el borde de la imagen, donde la información contextual disponible para el codificador es limitada; y (iii) lesiones multilobulares o irregulares cuya extensión espacial se desvía de la morfología aproximadamente convexa que domina la distribución de entrenamiento de Kvasir-SEG.

VI-E. Limitaciones y Consideraciones Clínicas

Varias limitaciones importantes restringen la generalización de los hallazgos presentes. En primer lugar, Kvasir-SEG fue recopilado en un único centro utilizando hardware de endoscopio específico, lo que limita las conclusiones que pueden extraerse sobre el rendimiento entre instituciones o

dispositivos. El cambio de dominio derivado de diferencias en la iluminación, los protocolos de cromoendoscopia, la configuración de compresión de vídeo o la calidad de la preparación mucosa entre centros clínicos podría degradar sustancialmente el rendimiento en el despliegue [30]. En segundo lugar, el conjunto de datos no incluye anotaciones explícitas de subtipo o morfología, lo que impide un análisis estratificado del rendimiento por categorías de pólipos (por ejemplo, tipos de la clasificación de París Ip, Is, IIa). En tercer lugar, la evaluación se realiza exclusivamente sobre fotogramas estáticos, mientras que la colonoscopia real implica secuencias de vídeo continuas donde la consistencia temporal y la latencia de inferencia por fotograma son clínicamente relevantes. En cuarto lugar, el tamaño del conjunto de datos de 1,000 imágenes, aunque estándar en la literatura, es modesto en relación con la plena diversidad de morfologías de pólipos encontradas en la práctica clínica.

VI-F. Comparación con el Estado del Arte: Perspectiva Crítica

Analizando métricas de otros artículos, podemos afirmar que el rendimiento de este modelo (Dice 0,9039, IoU 0,8427) es competitivo con las variantes de U-Net que utilizan codificadores similares o más pesados, y es consistente con el rango de rendimiento alcanzable por los métodos basados en CNN en Kvasir-SEG. Los métodos basados en transformers como Polyp-PVT [18] y TransFuse [12] reportan valores de Dice más altos en el mismo benchmark; sin embargo, estos modelos se benefician de recuentos de parámetros sustancialmente mayores y, en muchos casos, de procedimientos de entrenamiento más complejos. La U-Net estándar de Zenodo 2025 [37] reporta Dice 0,9449, pero la duración del entrenamiento (100 épocas frente a nuestras 33 épocas efectivas) y la definición exacta de la partición introducen factores de confusión. En conjunto, los resultados presentes confirman que una U-Net respaldada por ResNet34 entrenada con el protocolo propuesto alcanza un rendimiento cercano al estado del arte para su clase de arquitectura, demostrando un equilibrio efectivo entre complejidad del modelo y exactitud de segmentación.

VII. Conclusiones

Este artículo presentó un sistema de segmentación semántica basado en deep learning para la delineación de pólipos colorrectales en imágenes endoscópicas, construido sobre una arquitectura U-Net con un codificador ResNet34 preentrenado. El sistema fue entrenado y evaluado en el benchmark Kvasir-SEG y demostró un sólido rendimiento cuantitativo, alcanzando un coeficiente Dice medio de 0,9039 y un IoU de 0,8427 en un conjunto de prueba reservado de 100 muestras, junto con una Precisión, Recall, puntuación F2 y Exactitud a nivel de píxel competitivos. Estos resultados establecen la viabilidad de transferir representaciones preentrenadas en ImageNet al dominio de la colonoscopia dentro de un codificador-decodificador bien diseñado, optimizado con pérdida conjunta Dice-BCE, AdamW, cosine annealing y aumentación de datos.

El análisis cualitativo confirma que el modelo produce mapas de probabilidad bien calibrados con regiones de alta confianza espacialmente coherentes en la mayoría de los casos de prueba, mientras identifica la infrasegmentación de lesiones morfológicamente atípicas, de gran área o de bajo contraste como el principal modo de fallo. La tendencia de error conservadora (mayor Precisión que Recall) tiene implicaciones clínicas directas, motivando trabajos futuros sobre estrategias de ponderación de pérdidas que penalicen explícitamente los falsos negativos de forma más severa.

Varias direcciones para investigaciones futuras se desprenden de los hallazgos actuales. Arquitectónicamente, reemplazar el backbone ResNet34 por un codificador más potente como EfficientNet-B4 [26] o un Swin Transformer [27] puede mejorar la representación de casos difíciles sin un costo computacional prohibitivo. El entrenamiento en conjuntos de datos aumentados que combinen Kvasir-SEG con benchmarks adicionales como CVC-ClinicDB [28], ETIS-Larib [29] o CVC-ColonDB abordaría el sesgo de datos de un único centro y mejoraría la estabilidad al cambio de dominio. La integración de información temporal mediante extensiones recurrentes o convolucionales 3D ampliaría la aplicabilidad del análisis de fotogramas estáticos a la asistencia en colonoscopia de vídeo en tiempo real. Finalmente, incorporar una estimación de incertidumbre calibrada derivada del dropout de Monte Carlo o de conjuntos profundos en la interfaz clínica permitiría al sistema señalar predicciones de baja confianza para revisión manual obligatoria, abordando directamente los requisitos de seguridad del paciente exigidos por las vías regulatorias para la endoscopia asistida por IA.

Referencias

[1] F. Bray *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*, vol. 74, no. 3, pp. 229–263, 2024.

[2] International Agency for Research on Cancer (IARC), "Mexico Fact Sheet, GLOBOCAN 2022," *Global Cancer Observatory, World Health Organization*, 2024. [En línea]. Disponible en: https://gco.iarc.who.int/media/globocan/factsheets/populations/484-mexico-fact-sheet.pdf

[3] L. Tang *et al.*, "Advances in colorectal cancer screening: technological innovations, guideline discrepancies, and individualized strategies," *Frontiers in Oncology*, vol. 15, p. 1723546, 2025. DOI: 10.3389/fonc.2025.1723546.

[4] J. C. van Rijn *et al.*, "Polyp miss rate determined by tandem colonoscopy: a systematic review," *The American Journal of Gastroenterology*, vol. 101, no. 2, pp. 343–350, 2006.

[5] D. Jha *et al.*, "Kvasir-SEG: A Segmented Polyp Dataset," in *MultiMedia Modeling (MMM)*, Lecture Notes in Computer Science, vol. 11962, pp. 451–462, 2020.

[6] D. Jha *et al.*, "Kvasir-SEG: A Segmented Polyp Dataset," in *MultiMedia Modeling (MMM)*, Lecture Notes in Computer Science, vol. 11962, pp. 451–462, 2020.

[7] D. Jha *et al.*, "Kvasir-SEG: A Segmented Polyp Dataset," in *MultiMedia Modeling (MMM)*, Lecture Notes in Computer Science, vol. 11962, pp. 451–462, 2020.

[8] K. He, X. Zhang, S. Ren, y J. Sun, "Deep Residual Learning for Image Recognition," en *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[9] K. He, X. Zhang, S. Ren, y J. Sun, "Deep Residual Learning for Image Recognition," en *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[10] D. Jha *et al.*, "Automatic Polyp Segmentation using U-Net-ResNet50," en *ACM Multimedia Workshop*, 2021. arXiv:2012.15247.

[11] D.-P. Fan *et al.*, "PraNet: Parallel Reverse Attention Network for Polyp Segmentation," en *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2020. arXiv:2006.11392.

[12] Y. Zhang *et al.*, "TransFuse: Fusing Transformers and CNNs for Medical Image Segmentation," en *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2021. arXiv:2102.08005.

[13] F. Tang *et al.*, "Colorectal Polyp Segmentation Based on Deep Learning Methods: A Systematic Review," *Journal of Imaging Informatics in Medicine*, vol. 11, no. 9, 2025.

[14] O. Ronneberger, P. Fischer, y T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," en *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, vol. 9351, pp. 234–241, 2015. arXiv:1505.04597.

[15] O. Ronneberger, P. Fischer, y T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," en *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, vol. 9351, pp. 234–241, 2015. arXiv:1505.04597.

[16] Z. Zhou *et al.*, "UNet++: A Nested U-Net Architecture for Medical Image Segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pp. 3–11, 2018.

[17] D. Jha *et al.*, "ResUNet++: An Advanced Architecture for Medical Image Segmentation," en *IEEE International Symposium on Multimedia*, 2021.

[18] B. Dong *et al.*, "Polyp-PVT: Polyp Segmentation with Pyramid Vision Transformers," arXiv preprint arXiv:2108.06932, 2021.

[19] E. Asare *et al.*, "Explainable Polyp Segmentation using GradCAM and U-Net with ResNet34 Backbone," *Scientific African*, vol. 21, e01727, 2023.

[20] M. Sandler *et al.*, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," en *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4510–4520, 2018.

[21] D. P. Kingma y J. Ba, "Adam: A Method for Stochastic Optimization," en *International Conference on Learning Representations (ICLR)*, 2015.

[22] I. Loshchilov y F. Hutter, "Decoupled Weight Decay Regularization," en *International Conference on Learning Representations (ICLR)*, 2019.

[23] I. Loshchilov y F. Hutter, "SGDR: Stochastic Gradient Descent with Warm Restarts," en *International Conference on Learning Representations (ICLR)*, 2017.

[24] A. Buslaev *et al.*, "Albumentations: Fast and Flexible Image Augmentations," *Information*, vol. 11, no. 2, p. 125, 2020.

[25] C. Shorten y T. M. Khoshgoftaar, "A Survey on Image Data Augmentation for Deep Learning," *Journal of Big Data*, vol. 6, no. 1, p. 60, 2019.

[26] M. Tan y Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," en *International Conference on Machine Learning (ICML)*, 2019.

[27] Z. Liu *et al.*, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," en *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[28] J. Bernal *et al.*, "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.

[29] J. Silva *et al.*, "Toward Embedded Detection of Polyps in WCE Images for Early Diagnosis of Colorectal Cancer," *International Journal of Computer Assisted Radiology and Surgery*, vol. 9, no. 2, pp. 283–293, 2014.

[30] N. Tajbakhsh *et al.*, "Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning?," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1299–1312, 2016.

[31] F. Milletari, N. Navab y S.-A. Ahmadi, "V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation," en *Fourth International Conference on 3D Vision (3DV)*, pp. 565–571, 2016.

[32] I. Goodfellow, Y. Bengio y A. Courville, *Deep Learning*. MIT Press, 2016.

[33] J. Deng *et al.*, "ImageNet: A Large-Scale Hierarchical Image Database," en *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255, 2009.

[34] O. Russakovsky *et al.*, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[35] S. Kornblith *et al.*, "Do Better ImageNet Models Transfer Better?," en *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2661–2671, 2019.

[36] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," arXiv preprint arXiv:2010.11929, 2021.

[37] Zenodo, "U-Net Baseline for Polyp Segmentation," Zenodo Repository, 2025.