**Logotipo

Descripción generada automáticamente**

**Segmentación de imágenes de pólipos gastrointestinales para el diagnóstico temprano de cáncer colorrectal a través de Deep Learning**

Natalia López Grisales

Monografía presentada para optar al título de Especialista en Analítica y Ciencia de Datos

Asesor  
Nombres completos, Título académico más alto

Universidad de Antioquia  
Facultad de Ingeniería

Especialización en Analítica y Ciencia de Datos

Medellín, Antioquia, Colombia

2023

|  |  |
| --- | --- |
| **Cita** | (López Grisales, 2023) |
| **Referencia**  **Estilo APA 7 (2020)** | López Grisales, N. (2023). *Segmentación de imágenes de pólipos gastrointestinales para el diagnóstico temprano de cáncer colorrectal a través de Deep Learning* [Trabajo de grado especialización]. Universidad de Antioquia, Medellín, Colombia. |

**** 

Especialización en Analítica y Ciencia de Datos, CohorteVI.

Centro de Investigación Ambientales y de Ingeniería (CIA).

|  |  |
| --- | --- |
|  | Diagrama  Descripción generada automáticamente con confianza media |

Centro de Documentación Ingeniería (CENDOI)

**Repositorio Institucional:** http://bibliotecadigital.udea.edu.co

Universidad de Antioquia - www.udea.edu.co

Rector: John Jairo Arboleda Céspedes.

Decano: Julio Cesar Saldarriaga Molina

Jefe departamento: Diego José Luis Botia Valderrama

El contenido de esta obra corresponde al derecho de expresión de los autores y no compromete el pensamiento institucional de la Universidad de Antioquia ni desata su responsabilidad frente a terceros. Los autores asumen la responsabilidad por los derechos de autor y conexos.

**Tabla de contenido**

[Resumen 6](#_Toc151242397)

[Abstract 7](#_Toc151242398)

[1. Descripción del problema 8](#_Toc151242399)

[1.1. Problema de negocio 8](#_Toc151242400)

[1.2. Aproximación desde la analítica de datos 10](#_Toc151242401)

[1.3. Origen de los datos 11](#_Toc151242402)

[1.4. Métricas de desempeño 11](#_Toc151242403)

[2. Objetivos 13](#_Toc151242404)

[2.1. Objetivo general 13](#_Toc151242405)

[2.2. Objetivos específicos 13](#_Toc151242406)

[3. Datos 14](#_Toc151242407)

[3.1. Datos originales 14](#_Toc151242408)

[3.2. Datsets 14](#_Toc151242409)

[3.3. Analítica descriptiva 14](#_Toc151242410)

[5. References 16](#_Toc151242411)

**Lista de figuras**

[Fig 1. Imágenes de pólipos obtenidas durante colonoscopias (del dataset Kvasir-v2 dataset) 9](#_Toc151242377)

[Fig 2 Ejemplo de imagen contenida en el dataset 15](#_Toc151242378)

[Fig 3 Ejemplo de imagen contenida en el dataset 15](#_Toc151242379)

**Siglas, acrónimos y abreviaturas**

DAC Diagnóstico Asistido por Computador

**CCR** Cáncer colorrectal

**OMS** Organización Mundial de la Salud

ROC Receiver Operating Characteristic

IOU Interection over Union

ADR Adenoma Detection Rate

# Resumen

La identificación de pólipos en el colon y el recto permite un diagnóstico temprano del cáncer colorrectal, la segunda causa de muertes relacionadas con el cáncer en general en Estados Unidos. Esta identificación se hace mediante una colonoscopia, un procedimiento en el cual un profesional de la salud analiza las paredes del intestino grueso mediante imágenes transmitidas por una pequeña cámara introducida por el recto. Sin embargo, debido a variables como el tamaño y morfología de los pólipos, el entorno y mucosidades dentro del intestino, la duración del procedimiento, etc., algunos pólipos pueden pasar desapercibidos durante la colonoscopia, favoreciendo la posibilidad de desarrollar cáncer colorrectal ya que un pólipo maligno es un carcinoma temprano.

Para reducir la tasa de fallo en la identificación de pólipos gastrointestinales, se propone el desarrollo de un sistema de Diagnóstico Asistido por Computador (DAC) que permita identificar pólipos gastrointentinales mediante un algoritmo de segmentación de imágenes, usando imágenes médicas de colonoscopias. Para el entrenamiento del modelo de segmentación se usarán imágenes del dataset Kvasir-SEG, un conjunto de datos de acceso abierto de imágenes de pólipos gastrointestinales y las correspondientes máscaras de segmentación, anotadas manualmente y verificadas por un gastroenterólogo experimentado.

*Palabras clave*: pólipo, cáncer colorrectal, colonoscopia, segmentación de imágenes, Deep Learning.

# Abstract

Identification of polyps in the colon and rectum allows for early diagnosis of colorectal cancer, the second leading cause of cancer-related deaths overall in the United States. This identification is done by colonoscopy, a procedure in which a health professional examines the walls of the large intestine using images transmitted by a small camera inserted through the rectum. However, due to variables such as the size and morphology of the polyps, the environment and mucosities within the bowel, the duration of the procedure, etc., some polyps may go undetected during colonoscopy, favoring the possibility of developing colorectal cancer since a malignant polyp is an early carcinoma.

To reduce the failure rate in the identification of gastrointestinal polyps, we propose the development of a Computer Aided Diagnosis (CAD) system to identify gastrointestinal polyps by means of an image segmentation algorithm using medical colonoscopy images. Images from the Kvasir-SEG dataset, an open access dataset of gastrointestinal polyp images and the corresponding segmentation masks, manually annotated and verified by an experienced gastroenterologist, will be used to train the segmentation model.

*Keywords: polyp, colorectal cancer, colonoscopy, image segmentation, Deep Learning.*

# Descripción del problema

La identificación de pólipos en el colon y el recto permite un diagnóstico temprano del cáncer colorrectal, el tercer cáncer más comúnmente diagnosticado en Estados Unidos (Siegel, Wagle, Cercek, Smith, & Jemal, 2023). Esta identificación se hace mediante una colonoscopia, un procedimiento en el cual un profesional de la salud analiza las paredes del intestino grueso mediante imágenes transmitidas por una pequeña cámara introducida por el recto (American Cancer Society, 2023).

A pesar de la experiencia y las habilidades del profesional de la salud, algunos pólipos pueden pasar desapercibidos durante la colonoscopia por múltiples razones, como el tamaño y morfología de los pólipos, el entorno y mucosidades dentro del intestino, la duración del procedimiento, el estado de agotamiento del profesional etc. (Ahmad, y otros, 2019).

El diagnóstico asistido por computador ofrece una interesante solución para reducir el porcentaje de fallos en la identificación de pólipos en colonoscopias (Ahmad, y otros, 2019), por esto se busca desarrollar un modelo de machine learning que permita mejorar la identificación de pólipos en las colonoscopias para una detección temprana y precisa del cáncer colorrectal.

## Problema de negocio

El cáncer colorrectal (CCR) es el tercer cáncer más comúnmente diagnosticado y la tercera causa más común de muerte relacionada con el cáncer en los Estados Unidos. Sin embargo, ocupa el segundo lugar en muertes relacionadas con el cáncer en general y es la principal causa en hombres menores de 50 años (Siegel, Wagle, Cercek, Smith, & Jemal, 2023).

En términos de incidencia y mortalidad, el CCR obtuvo el tercer mayor número de casos de cáncer y el segundo mayor número de muertes por cáncer atribuibles a factores de riesgo, evaluados tanto en hombres como en mujeres según un estudio realizado en Estados Unidos en 2018 (Islami, y otros, 2018). En este estudio, también se enfatiza en la necesidad de una implementación generalizada de acceso a la atención preventiva de la salud y la concienciación sobre los factores de riesgo que podrían haberse mitigado mediante estrategias preventivas eficaces.

Según la Organización Mundial de la Salud (OMS), los cambios en el estilo de vida y las pruebas de detección periódicas pueden ayudar a prevenir el cáncer colorrectal (World Health Organization, 2023). Dentro de los métodos de diagnóstico para el cáncer colorrectal se encuentran: el examen físico, las imágenes (como la ecografía abdominal, la tomografía computarizada y la resonancia magnética), el examen del interior del colon mediante colonoscopia o sigmoidoscopia, etc. (World Health Organization, 2023).

La realización de una colonoscopia de detección se asocia con una reducción de la mortalidad por cáncer de colon proximal y distal (Nishihara, y otros, 2013). En comparación con la ausencia de exámenes de detección endoscópicos, se presentó un riesgo 67 % menor de muerte por cáncer colorrectal en general después de realizar una colonoscopia de detección (Doubeni, y otros, 2018). La colonoscopia es un examen por imágenes mínimamente invasivo de todo el colon y el recto que permite la detección de pólipos, la caracterización de la densidad y el sitio del tumor (Conteduca, Sansonno, Russi, & Dammacco, 2013). En este procedimiento, un médico utiliza un tubo flexible con una cámara en un extremo, llamado colonoscopio o endoscopio, para observar el interior del recto y el colon en busca de pólipos (American Cancer Society, 2023).

Un pólipo colorrectal maligno se refiere a un adenoma (protrusión anormal de la mucosa: **Error! Reference source not found.**) que parece benigno macroscópicamente, pero en el que hay invasión a través de la mucosa muscular hacia la submucosa, por lo tanto, un pólipo maligno es un carcinoma temprano (Williams, y otros, 2013).

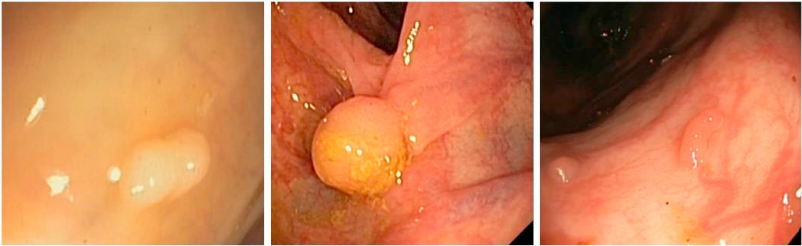


Fig 1. Imágenes de pólipos obtenidas durante colonoscopias (del dataset Kvasir-v2 dataset)

La colonoscopia es un método imperfecto y de calidad variable, con tasas de error del 17 % al 28% para pólipos de cualquier tamaño, del 2,1 % al 12 % para pólipos grandes y del 5 % al 11 % para adenomas avanzados. Además, las tasas de detección de adenomas y lesiones serradas sésiles varían entre colonoscopistas (Park & Cha, 2022). La evidencia indirecta sugiere que, a pesar de la experiencia y las habilidades de los médicos, durante una colonoscopia larga, algunos pólipos pueden pasar desapercibidos, incluso cuando están dentro del campo de visión endoscópico, debido a su tamaño diminuto o porque están ocultos detrás de los pliegues intestinales, por lo cual los endoscopistas podrían no identificar las lesiones (Ahmad, y otros, 2019). Los pólipos serrados sésiles (lesiones que luego progresan a CCR) suelen ser planas y estar cubiertas por una capa mucosa, lo que significa que pueden ser difíciles de detectar en la colonoscopia (Shine, Bui, & Burgess, 2020).

Debido a las múltiples formas y tamaños de los pólipos y la dificultad para ser identificados al estar cubiertos de heces, mucosa y otras estructuras circundantes, es esencial crear un sistema de Diagnóstico Asistido por Computador (DAC) para detectar los pólipos (Tomar, 2021).

## Aproximación desde la analítica de datos

Para superar las limitaciones mencionadas en la sección anterior, el DAC es una alternativa atractiva, ya que podría ayudar a mejorar la tasa de detección y evitar los errores en la caracterización de los pólipos (Sánchez-Montes, Bernal, García-Rodríguez, Córdova, & Fernández-Esparrach, 2020).

A través del tiempo se han desarrollado múltiples estrategias para mejorar la calidad de las colonoscopias, optimizando la identificación y caracterización de lesiones. Estos avances van desde mejoras a dispositivos actuales, como acoplamiento de nuevas herramientas y desarrollo de sistemas computacionales de soporte (Sánchez-Montes, García-Rodríguez, Córdova, Pellisé, & Fernández-Esparrach, 2020).

Mediante contenido de videos e imágenes de colonoscopias y la participación de profesionales que analicen y caractericen el contenido, es posible la recolección, anotación y publicación de bases de datos para la posterior creación de propuestas para la detección automática de pólipos (Sánchez-Montes, Bernal, García-Rodríguez, Córdova, & Fernández-Esparrach, 2020).

Actualmente, se han propuesto diferentes sistemas de DAC para la colonoscopia usando procesamiento digital de imágenes, algoritmos de machine learning, deep learning, etc. También se ha avanzado en la publicación de conjuntos de datos que permitan seguir explorando este campo de la medicina (Hsu, y otros, 2021).

Se pretende desarrollar un modelo de machine learning, usando un conjunto de datos público, que permita mejorar la identificación de pólipos en las colonoscopias para una detección temprana y precisa del CCR.

## Origen de los datos

Kvasir SEG es un conjunto de datos de acceso abierto que contiene imágenes del interior del intestino grueso, obtenidas durante colonoscopias, y las correspondientes máscaras de segmentación, anotadas manualmente y verificadas por un gastroenterólogo experimentado (Jha, et al., 2020).

Los datos se recogen utilizando equipos endoscópicos de Vestre Viken Health Trust (VV) en Noruega. La VV consta de 4 hospitales y brinda atención médica a 470.000 personas. Uno de estos hospitales (el Hospital Bærum) cuenta con un gran departamento de gastroenterología desde el que se continúan adquiriendo datos, lo que hará que el conjunto de datos sea más amplio en el futuro. Además, las imágenes son cuidadosamente anotadas por uno o más expertos médicos de VV y el Registro de Cáncer de Noruega (CRN). El CRN proporciona nuevos conocimientos sobre el cáncer a través de la investigación, forma parte de la Autoridad Regional de Salud del Sudeste de Noruega y está organizada como una institución independiente bajo el Fideicomiso del Hospital Universitario de Oslo. El CRN es responsable de los programas nacionales de cribado del cáncer con el objetivo de prevenir la muerte por cáncer mediante el descubrimiento de cánceres o lesiones precancerosas lo antes posible (Jha, et al., 2020).

## Métricas de desempeño

Los modelos se evaluarán usando métricas de machine learning cómo precision, recall, accuracy, f1 score y el área bajo la curva ROC (Receiver Operating Characteristic). Para el análisis de sensibilidad y especificidad se definen los siguientes términos:

* Verdadero positivo (TP) El número de muestras correctamente identificadas. El número de fotogramas con un hallazgo endoscópico que se identifica correctamente como un fotograma con un hallazgo endoscópico.
* Verdadero negativo (TN) El número de muestras negativas correctamente identificadas, es decir, marcos sin un hallazgo endoscópico que se identifica correctamente como un marco sin un hallazgo endoscópico.
* Falso positivo (FP) El número de muestras erróneamente identificadas, es decir, una comúnmente llamada "falsa alarma". Marcos sin un hallazgo endoscópico que se identifica erróneamente como un marco con un hallazgo endoscópico.
* Falso negativo (FN) El número de muestras negativas identificadas erróneamente. Marcos sin un hallazgo endoscópico que erróneamente se identifica como un marco con un hallazgo endoscópico.

Para la tarea de segmentación de imágenes médicas, las métricas que se usarán son:

* Coeficiente de Dice
* Intersección sobre la unión (IOU)

Las métricas de negocio a usar son:

* Cantidad de pólipos identificados
* Tamaño del pólipo
* Tasa de detección de adenoma, ADR por sus siglas en inglés. (número de pacientes con uno o más adenomas dividido por el número total de pacientes que se sometieron a una colonoscopia de detección)

Se espera obtener una tasa de detección de adenomas superior al 30 %, mediante modelos de machine learning con una precisión superior al 85%. El Grupo de Trabajo sobre Calidad en Endoscopia de la Sociedad Americana de Endoscopia Gastrointestinal (ASGE)/Colegio Americano de Gastroenterología (ACG) recomienda alcanzar para la colonoscopia de detección un ADR de al menos 30% para los hombres y 20% para las mujeres (25% para todos los pacientes) de ≥50 años (Park & Cha, 2022).

# Objetivos

## Objetivo general

Desarrollar un sistema de Diagnóstico Asistido por Computador (DAC) que permita identificar pólipos gastrointestinales ubicados en el colon o en el recto usando imágenes médicas obtenidas de colonoscopias.

## Objetivos específicos

Preprocesar digitalmente las imágenes de colonoscopias del conjunto de datos Kvasir SEG como preparación para su uso en modelos de machine learning.

Desarrollar, entrenar y evaluar un algoritmo de segmentación de imágenes usando una técnica de Deep learning para la identificación y segmentación de pólipos gastrointestinales en las imágenes de las colonoscopias.

Diseñar un método para presentar los pólipos identificados por el modelo sobre las imágenes de colonoscopias, como sistema computacional de soporte para la visualización.

# Datos

## Datos originales

El conjunto de datos Kvasir-SEG contiene 1000 imágenes de pólipos y su correspondiente ground truth. La resolución de las imágenes varía de 332x487 a 1920x1072 píxeles. Los archivos de imagen se codifican mediante compresión JPEG.

El cuadro delimitador (puntos de coordenadas) de las imágenes correspondientes se almacena en un archivo JSON.

El conjunto de datos tiene un tamaño de 50.1 MB distribuido en una carpeta con las imágenes con un tamaño de 37.7 MB, una carpeta con el ground truth llamada mask con tamaño 12.2 MB y un archivo JSON con los cuadros delimitadores de tamaño 144KB.

El conjunto de datos de acceso abierto se puede descargar fácilmente con fines educativos y de investigación a través de la página web [Conjuntos de datos de Simula - Kvasir SEG](https://datasets.simula.no/kvasir-seg/).

## Datsets

Para construir los datasets de entrenamiento y validación inicialmente se lee el conjunto de datos, incluyendo imágenes y archivos con las etiquetas, luego se realiza un preprocesamiento de imágenes donde se aplican transformaciones buscando estandarizar y resaltar los pixeles de interés.

Las trasformaciones consisten en:

* Redimensionamiento: El tamaño de la imagen de entrada se estandariza a 320x320.
* Filtrado: se aplica el filtro bilateralFilter para reducir el ruido no deseado y conservar los bordes nítidos.
* Normalizado: se usa la técnica Min-Max para normalizar los valores de cada pixel.

Posteriormente, después de las transformaciones, se extraen las características y la etiqueta de imagen. Finalmente se construyen los sets de entrenamiento y prueba: X\_train, X\_test, y\_train, y\_test.

## Analítica descriptiva

Algunos ejemplos de las imágenes contenidas en el dataset:



Fig 2 Ejemplo de imagen contenida en el dataset



Fig 3 Ejemplo de imagen contenida en el dataset

# References

Ahmad, O. F., Soares, A. S., Mazomenos, E., Brandao, P., Vega, R., Seward, E., . . . Lovat, L. B. (2019). Artificial intelligence and computer-aided diagnosis in colonoscopy: current evidence and future directions. *The lancet. Gastroenterology & hepatology, 4*(1), 71-80. doi:10.1016/s2468-1253(18)30282-6

American Cancer Society. (2023, 10 3). *Colonoscopy*. Retrieved from Cancer.org: https://www.cancer.org/cancer/diagnosis-staging/tests/endoscopy/colonoscopy.html

Conteduca, V., Sansonno, D., Russi, S., & Dammacco, F. (2013). Precancerous colorectal lesions. *International journal of oncology, 43*(4), 973-984. doi:10.3892/ijo.2013.2041

Doubeni, C. A., Corley, D. A., Quinn, V. P., Jensen, C. D., Zauber, A. G., Goodman, M., . . . Fle. (2018). Effectiveness of screening colonoscopy in reducing the risk of death from right and left colon cancer: a large community-based study. *Gut, 67*(2), 291-298. doi:10.1136/gutjnl-2016-312712

Hsu, C.-M., Hsu, C.-C., Hsu, Z.-M., Shih, F.-Y., Chang, M.-L., & Chen, T.-H. (2021). Colorectal polyp image detection and classification through grayscale images and deep learning. *Sensors (Basel, Switzerland), 21*(1/), 5995. doi:10.3390/s21185995

Islami, F., Sauer, A. G., Miller, K. D., Siegel, R. L., Fedewa, S. A., Jacobs, E. J., . . . Jemal, A. (2018). Proportion and number of cancer cases and deaths attributable to potentially modifiable risk factors in the United States. *CA: a cancer journal for clinicians, 68*(1), 31-54. doi:10.3322/caac.21440

Jha, D., Smedsrud, P. H., Riegler, M. A., Halvorsen, P., de Lange, T., Johansen, D., & Johansen, H. D. (2020). Kvasir-SEG: A Segmented Polyp Dataset. *MultiMedia Modeling*. Springer International Publishing. doi:10.1007/978-3-030-37734-2\_37

Nishihara, R., Wu, K., Lochhead, P., Morikawa, T., Liao, X., Qian, Z. R., . . . Ogino, S. (2013). Long-term colorectal-cancer incidence and mortality after lower endoscopy. *The New England journal of medicine, 369*(12), 1095-1105. doi:10.1056/nejmoa1301969

Park, S. B., & Cha, J. M. (2022). Quality indicators in colonoscopy: the chasm between ideal and reality. *Clinical endoscopy, 55*(3), 332-338. doi:10.5946/ce.2022.037

Sánchez-Montes, C., Bernal, J., García-Rodríguez, A., Córdova, H., & Fernández-Esparrach, G. (2020). Revisión de métodos computacionales de detección y clasificación de pólipos en imagen de colonoscopia. *Gastroenterologia y hepatologia, 43*(4), 222-232. doi:10.1016/j.gastrohep.2019.11.004

Sánchez-Montes, C., García-Rodríguez, A., Córdova, H., Pellisé, M., & Fernández-Esparrach, G. (2020). Tecnologías de endoscopia avanzada para mejorar la detección y caracterización de los pólipos colorrectales. *Gastroenterologia y hepatologia, 43*(1), 46-56. doi:10.1016/j.gastrohep.2019.09.008

Shine, R., Bui, A., & Burgess, A. (2020). Quality indicators in colonoscopy: an evolving paradigm: Quality indicators in colonoscopy. *ANZ journal of surgery, 90*(3), 215-221. doi:10.1111/ans.15775

Siegel, R. L., Wagle, N. S., Cercek, A., Smith, R. A., & Jemal, A. (2023). Colorectal cancer statistics, 2023. *CA: a cancer journal for clinicians, 73*(3), 233-254. doi:10.3322/caac.21772

Tomar, N. K. (2021, 01 11). *Automatic polyp segmentation using fully convolutional neural network.* Retrieved from arXiv [eess.IV]: http://arxiv.org/abs/2101.04001https://arxiv.org/abs/2101.04001

Williams, J. G., Pullan, R. D., Hill, J., Horgan, P. G., Salmo, E., Buchanan, G. N., . . . Haboubi, N. (2013). Management of the malignant colorectal polyp: ACPGBI position statement. *Colorectal disease: the official journal of the Association of Coloproctology of Great Britain and Ireland, 15*(s2), 1-38. doi:10.1111/codi.12262

World Health Organization. (2023, 07 11). *Colorectal cancer*. Retrieved from Who.int: https://www.who.int/news-room/fact-sheets/detail/colorectal-cancer