UNIVERSIDAD DE EXTREMADURA

TESIS DOCTORAL

MEJORAS EN LA DETECCIÓN, LOCALIZACIÓN Y SEGMENTACIÓN AUTOMÁTICA DE PÓLIPOS COLORRECTALES BASADAS EN TÉCNICAS DE APRENDIZAJE PROFUNDO

Luisa Fernanda Sánchez Peralta

Programa de Doctorado en Tecnologías Informáticas

2021

UNIVERSIDAD DE EXTREMADURA

TESIS DOCTORAL

MEJORAS EN LA DETECCIÓN, LOCALIZACIÓN Y SEGMENTACIÓN AUTOMÁTICA DE PÓLIPOS COLORRECTALES BASADAS EN TÉCNICAS DE APRENDIZAJE PROFUNDO

Luisa Fernanda Sánchez Peralta

Conformidad de los directores:

Fdo. Dr. D. J. Blas Pagador Carrasco

Fdo. Dr. D. Artzai Picón Ruiz

La conformidad de los directores de la tesis consta en el original en papel de esta Tesis Doctoral

Programa de Doctorado en Tecnologías Informáticas

2021

Agradecimientos institucionales

Esta tesis doctoral ha sido financiada a través del proyecto PICCOLO, financiado por el Programa de Innovación e Investigación Horizon2020 de la Unión Europea, bajo el acuerdo de financiación Nº 732111, y por la ayuda GR18199 a Grupos de Investigación Catalogados, de la Consejería de Economía, Ciencia y Agenda Digital de la Junta de Extremadura y cofinanciada por el Fondo Europeo de Desarrollo Regional (FEDER).

La realización de esta tesis se ha llevado a cabo en el Centro de Cirugía de Mínima Invasión Jesús Usón y en la ICTS NANBIOSIS.

A mis padres, Piru y Luis

Agradecimientos personales

A mis directores, el Dr. Blas Pagador y el Dr. Artzai Picón, por el tiempo y esfuerzo dedicado a que esta tesis llegara a buen puerto. Blas, gracias por apostar y confiar en mi, no solo con la tesis sino desde el primer día en el que me firmaste los papeles de la beca de tecnólogo. Ha sido un viaje de muchos años, por lo que ha dado para buenos y malos momentos de los que siempre he aprendido. Artzai, gracias por subirte al carro sin ni siquiera conocerme y por transmitirme el sabio consejo del pequeño Xabier de que train y test no se mezclan. Gracias a los dos de corazón por hacer de esta tesis una realidad, espero que sigamos dando pasos juntos para que pueda seguir creciendo personal y profesionalmente con vosotros.

Al Centro de Cirugía de Mínima Invasión Jesús Usón y al Dr. Francisco M. Sánchez Margallo, por posibilitar la realización de esta tesis dentro del proyecto PICCOLO, donde he tenido la ocasión de trabajar con un equipo con una calidad humana inigualable. A los coautores de las contribuciones de esta tesis, porque sin su apoyo y trabajo esta hubiera sido imposible. Me gustaría mencionar expresamente a Virginia, Águeda, Ben, Elena, Alfonso, Jorge, Juanfran y Luis Bote, y sobre todo a Cristina, no sólo por lidiar para sacar PICCOLO adelante, sino también por preocuparte e involucrarte en cualquier problema que ha ido surgiendo durante el desarrollo de esta tesis. Estoy convencida que en no mucho podrás estar escribiendo tú también los agradecimientos de tu tesis.

A la gente que habita el CCMIJU, porque es lo más valioso que tiene el centro y porque no sólo sois unos grandísimos profesionales, sino mejores personas. A Ferni, no solo por todo y cada uno de los artículos conseguidos, sino por todos esos ratitos de charla en cualquier sitio y momento sobre cualquier cosa. Por preocuparte, por estar siempre. Pero sin duda, por lo mejor, por darme un tocayo de cumpleaños. A Fede, por facilitarme el acceso a los clínicos, tan importantes en esta tesis, y por todos los consejos desde la experiencia para ir lidiando con lo que es e implica hacer una tesis. A Eva, por tener siempre una sonrisa y unos ánimos inagotables, por estar al pie del cañón en todo momento. A Pilar, por preocuparse por mi casi tanto como mi madre. A JL Antequera, por estar siempre disponible para evitar que muera congelada. A Pepe y Carlos, por las conversaciones de los desayunos en las que no nos entienden. Al club del táper, aunque lo tengamos en standby, por hacer el rato de la comida muchísimo más entretenido. A Julia y Federica, por resolver todas y cada una de las mil preguntas que surgen durante todo el papeleo que conlleva el depósito de la tesis. Al resto de compañeros que están o han pasado por el CCMIJU, Iván, Javi terapia, Alejandro, Bea Macías, Alberto, Jose micro, Marina, Blanca, Silvia, Javier Sánchez, Claudia, Virginia, Carmen, Sonia, Anabel..., porque todos siempre os habeis interesado por cómo iba con la tesis, y habéis tenido siempre consejos y palabras de ánimo cuando lo he necesitado.

A los bioinformáticos, por ser la familia del trabajo. A Juanmi y David, por tener siempre la solución a cada problema que os planteo, porque sé que en el fondo os molan los retos. A Nuria y María, por las mil y una modificaciones a cualquiera de las cosas que hemos preparado. Nuri, gracias por ponerle portada a este libro de tesis e implicarte como lo has hecho. A Joaquín, por las charlas, las risas y los chascarrillos, por los ánimos y los consejos cuando se ve todo oscuro. A las nuevas incorporaciones en bioingeniería, David, Víctor y María, y a los que ya han volado, Raúl Bojo, Luis Bote, JL Moyano, Marina, Paco Esteban y Enrique, porque cada cual a vuestra manera

habeis aportado algo durante estos cuatro años de tesis. Por último, especialmente a Juanfran, por estar siempre dispuesto a echar una mano y por ese apoyo incondicional, sobre todo en los últimos meses, que ha hecho que saquemos todo para adelante. Te debo mucho, juanfransciscortega, así que mil gracias. A todos, gracias por hacer el día a día en el ala oeste del centro mucho mejor.

A Alejandro Moreo, por estar en los principios del deep y porque no podré ver una ReLU sin brindar contigo, aunque sea mentalmente.

A Ana González, por pasar de project coordinator a amiga, gracias por hacer que las reuniones de proyectos sean mejores y por preocuparte por mi tesis siempre aunque sea desde la distancia.

A los fitweleros MJ y David, por preocuparse de mi salud física y mental, aún en los últimos meses cuando no nos hemos podido ni ver. Espero que más pronto que tarde podamos volver a reunir al cuerpo de baile.

A las niñas del Buen Aire, en especial a Carmen y María, porque aunque nos veamos cada vez más de cuando en cuando, retomamos como si nos hubiéramos visto ayer.

A Montaña, por los paseos, las terapias y por estar siempre, independientemente del tiempo que pasemos sin vernos. Si alguien sabe el esfuerzo que supone una tesis, esa eres tú.

A mis amigos del poblado, en especial a Rocío, Mercedes y Jose, por esas risas infinitas, por muchas más cenas en el Chaplin, por más partidas al Catán. Es todo un lujo teneros en mi vida y espero que sigamos disfrutando del gastroturismo mucho más.

A mis padres y a mi hermano, por ser los tres pilares esenciales en mi vida. Mamá, papá, gracias porque soy lo que soy por vosotros, por vuestro ejemplo, vuestro esfuerzo y vuestro amor sin límites, porque mis logros son vuestros. Tete, gracias por estar siempre ahí y por demostrarme que con empeño, dedicación y esfuerzo se consigue todo. Espero poder devolveros todo el tiempo que el confinamiento y la tesis me ha hecho robaros.

A Juan Carlos, por llegar, por liarme, por cuidarme y entenderme aún cuando no hablo, por hacer tuya esta tesis, por ser el cuarto pilar. No veo el momento en que podamos disfrutar sin tener la palabra tesis sobrevolando. A por los viajes, sueños y líos que están por venir, que estoy segura de que no van a ser pocos.

Resumen

El cáncer colorrectal es una de las principales causas de muerte en los países desarrollados: una de cada diez personas lo sufrirán a lo largo de su vida y se estima que causa unas 700.000 fallecimientos anualmente. La detección precoz, mediante los programas de cribado, es la principal estrategia para reducir la incidencia del cáncer colorrectal, ya que la tasa de supervivencia aumenta si el cáncer se detecta en fase inicial, pasando del 18 % de supervivencia a los 5 años al 88,5 %. La colonoscopia es la técnica diagnóstica habitualmente utilizada para la detección y tratamiento de pólipos colorrectales, lesiones precursoras del cáncer colorrectal.

En los últimos años, los sistemas de inteligencia artificial basados en aprendizaje profundo (o *deep learning*) se han aplicado con gran éxito en multitud de tareas relacionadas con la visión por ordenador en las que se cuenta con volúmenes ingentes de datos: clasificación, segmentación o anotación de imágenes naturales, identificación de objetos o recuperación de la información. De igual manera, el aprendizaje profundo ha comenzado a aplicarse en el ámbito médico, donde ha llegado a obtener resultados comparables a los obtenidos por expertos clínicos. No obstante, presenta un problema general que radica en la cantidad limitada de información disponible, ya que la generación del conjunto de datos requiere un conocimiento clínico experto de facultativos que habitualmente disponen de un tiempo muy reducido para estas tareas.

El objetivo principal de esta tesis es mejorar la detección, localización y segmentación automática de pólipos colorrectales aplicando tecnologías de aprendizaje profundo. Para ello, inicialmente se han analizado las necesidades clínicas y se ha realizado una revisión sistemática para contextualizar el estado del arte de dichos métodos. Posteriormente se ha realizado un análisis de diferentes transformaciones y rangos para la estrategia de aumento de datos y se ha propuesto una nueva función de pérdida, *eigenloss*, basada en un análisis de componentes principales. Se ha creado también un conjunto de datos de acceso abierto con 3.433 imágenes manualmente segmentadas y puestas a disposición de la comunidad científica, junto con una comparativa de diferentes modelos y conjuntos de datos.

Palabras clave: cáncer colorrectal, colonoscopia, pólipo, aprendizaje profundo, detección, localización, segmentación, aumento de datos, *eigenloss*, conjunto de datos.

Abstract

Colorectal cancer is one of the leading death causes in developed countries: one in ten people will suffer from it in their lifetime and it estimated to cause 700,000 deaths annually. Early detection, through screening programmes, is the main strategy to reduce the incidence of colorectal cancer, as the survival rate increases if the cancer is detected at an early stage, from 18 % 5-years survival rate to 88.5 %. Colonoscopy is the diagnostic technique most commonly used for the detection and treatment of colorectal polyps, precursor lesions of colorectal cancer.

In recent years, artificial intelligence systems based on deep learning have been applied with great success in several computer vision-related tasks involving huge data volumes: classification, segmentation or annotation of natural images, object identification or information retrieval. Similarly, deep learning has started to be applied in the medical field, where it has achieved results comparable to those obtained by clinical experts. However, it presents a general challenge that lies in the limited amount of information available, since the generation of the dataset requires expert clinical knowledge by practitioners who usually have little time available.

The main objective of this thesis is to improve the automatic detection, localisation and segmentation of colorectal polyps by applying deep learning technologies. To this end, a needs analysis and a systematic review have been carried out to contextualise the state of the art of these methods. Subsequently, an analysis of different transforms and ranks for the data augmentation strategy has been performed and a new loss function, *eigenloss*, based on principal component analysis has been proposed. An open access dataset has also been created including 3,433 manually segmented images and made available to the scientific community, together with a comparison of different models and datasets.

Keywords: Colorectal cancer, colonoscopy, polyp, deep learning, detection, localization, segmentation, data augmentation, *eigenloss*, dataset.

Índice general

Índice de figuras III

Índice de tablas V

Acrónimos y abreviaturas VII

Compendio de publicaciones IX

1. Justificación unitaria de la tesis 1

2. Introducción 3
2.1. Cáncer colorrectal . . . 3
2.1.1. Colonoscopias y detección temprana de lesiones . . . 5
2.1.2. Tipos de pólipos . . . 7
2.2. Inteligencia artificial y aprendizaje profundo en el procesado de imágenes médicas 9
2.3. Objetivos . . . 11

3. Técnicas de aprendizaje profundo aplicado a procesado de imagen 13
3.1. Aprendizaje profundo supervisado . . . 13
3.2. Arquitecturas . . . 16
3.2.1. Tipos de capas . . . 16
3.2.2. Redes neuronales convolucionales . . . 18
3.2.3. Redes neuronales completamente convolucionales . . . 19
3.3. Aumento de datos . . . 20
3.4. Función de pérdida . . . 21

4. Detección, localización y segmentación automática de pólipos colorrectales basadas en técnicas de aprendizaje profundo 23
4.1. Métodos de detección, localización y segmentación . . . 23
4.2. Limitaciones de los métodos actuales . . . 25
4.3. Conjuntos de datos de pólipos colorrectales . . . 26
4.3.1. Conjunto de datos públicos . . . 26
4.3.2. Conjunto de datos de PICCOLO . . . 26
4.4. Métricas . . . 27

5. Contribuciones de esta tesis 31
5.1. Necesidades médicas relacionadas con la tecnología endoscópica y la colonoscopia para diagnóstico de cáncer colorrectal . . . 31

5.2. Aprendizaje profundo para encontrar pólipos en colonoscopia: una revisión sistemática de la literatura . . . 33
5.2.1. Análisis de conjuntos de datos públicos de imágenes colorrectales . . . 33
5.2.2. Análisis de métodos basados en aprendizaje profundo para detección, localización y segmentación de pólipos . . . 33
5.2.3. Análisis de métricas empleadas . . . 37
5.2.4. Identificación de retos y recomendaciones para la aplicación de métodos de aprendizaje profundo para pólipos colorrectales . . . 37
5.3. Efecto de transformaciones para aumento de datos en segmentación de pólipos . . 37
5.4. Eigenloss: función de pérdida combinada basada en PCA para segmentación de pólipos . . . 40
5.5. Conjunto de datos de PICCOLO: comparativa de modelos y conjuntos de datos . 42

6. Conclusiones **47**

7. Bibliografía **49**

8. Resultados de esta tesis **61**
8.1. Medical needs related to the endoscopic technology and colonoscopy for colorectal cancer diagnosis . . . 65
8.2. Deep learning to find colorectal polyps in colonoscopy: a systematic literature review . . . 80
8.3. Unravelling the effect of data augmentation transformations in polyp segmentation 106
8.4. Eigenloss: combined PCA-based loss function for polyp segmentation . . . 122
8.5. PICCOLO white-light and narrow-band imaging colonoscopic dataset: a performance comparative of models and datasets . . . 144

A. Material suplementario **159**

Índice de figuras

2.1. Tasas de incidencia y mortalidad de CCR en Europa. 4
2.2. Detección precoz sintomática y cribado. 4
2.3. Imagenes extraídas de colonoscopias HD. 6
2.4. Tipos macroscópicos de carcinoma colorrectal avanzado. 7
2.5. Clasificación endoscópica de París para lesiones superficiales. 8
2.6. Clasificación NICE para pólipos colorrectales. 8
2.7. Evolución de la inteligencia artificial hacia el aprendizaje profundo. 9

3.1. Representación esquemática de una neurona biológica y otra artificial. 13
3.2. Aprendizaje supervisado. 15
3.3. Representación esquemática de los diferentes tipos de capas. 17
3.4. Arquitectura básica de una red neuronal convolucional (CNN). 18
3.5. Arquitectura básica de una red completamente convolucional (FCN). 20

4.1. Representación esquemática de las tareas consideradas en esta tesis. 24
4.2. Evolución temporal del número de publicaciones. 25
4.3. Ejemplos de las imágenes proporcionadas por CVC-EndoSceneStill y Kvasir-SEG. 26
4.4. Matriz de confusión. 27
4.5. Comparación de métricas. 29

5.1. Redes para cada una de las tareas. 34
5.2. Representación esquemática de los cinco enfoques para métodos de DL. 35
5.3. Categorización de trabajos por tarea 35
5.4. Imagen original e imágenes transformadas [138]. 38
5.5. Redes troncales y arquitecturas codificador-decodificador. 40
5.6. Ejemplos de las imágenes proporcionadas por PICCOLO. 44
5.7. Índice Jaccard para cada uno de los modelos. 46

Índice de tablas

5.1. Necesidades clínicas y del sistema CAD. 32
5.2. Ventajas y desventajas de las diferentes categorías de modelos de DL 36
5.3. Resultados de Jaccard para los diferentes modelos y funciones de pérdida. 41
5.4. Resultados de la PCA. 42
5.5. Comparación entre modelos entrenados con *eigenloss* y otros trabajos. 43
5.6. Distribución de imágenes en el conjunto de datos de PICCOLO. 45
5.7. Índice Jaccard para cada experimento. 46

Acrónimos y abreviaturas

ACC	Exactitud
ADR	*Adenoma Detection Rate* - Tasa de detección de adenoma
AI	*Artificial Intelligence* - Inteligencia artificial
ASR	*Age-Standardised Rate* - Tasa estandarizada por edad
BN	Capa de normalización por lotes
CAD	*Computer Assisted Detection and Diagnosis* - Detección y diagnóstico asistido por ordenador
CCR	Cáncer colorrectal
CNN	*Convolutional Neural Network* - Red neuronal convolucional
CONV	Capa convolucional
CPU	*Central Processing Unit* - Unidad de procesamiento central
DECONV	Capa de convolución transpuesta
DL	*Deep Learning* - Aprendizaje profundo
DROPOUT	Capa de abandono
ESGE	*European Society of Gastrointestinal Endoscopy* - Sociedad Europea de Endoscopia Gastrointestinal
FUSE	*Full Spectrum Endoscope* - Endoscopio de amplio espectro
FCN	*Fully Convolutional Network* - Red completamente convolucional
FC	Capa completamente conectada
FN	*False Negatives* - Falsos negativos
FP	*False Positives* - Falsos positivos
GAN	*Generative Adversarial Network* - Red generativa antagónica
GPU	*Graphic Processing Unit* - Unidad de procesamiento gráfico
HD	*High Definition* - Alta definición

ILSVRC	*ImageNet Large Scale Visual Recognition Competition*
IoU	*Intersection over Union* - Intersección sobre la unión
LDA	*Latent Dirichlet Allocation* - Asignación latente de Dirichlet
LSTM	*Long Short-Term Memory* - Redes de gran memoria de corto plazo
MAXPOOL	Capa de muestreo del valor máximo
MICCAI	*International Conference on Medical Image Computing and Computer Assisted Intervention*
NBI	*Narrow Band Imaging* - Imagen de banda estrecha
NICE	*NBI International Colorectal Endoscopic classification* - Clasificación endoscópica colorrectal internacional NBI
PCA	*Principal Component Analysis* - Análisis de componentes principales
POOL	Capa de muestreo
PREC	Precisión
ReLU	*Rectified Linear Unit* - Unidad lineal rectificada
REC	Exhaustividad
RGB	*Red Green Blue* - Rojo verde azul
RNN	*Recurret Neural Network* - Red neuronal recurrente
SPEC	Especificidad
SGD	*Stochastic Gradient Descent* - Descenso de gradiente estocástico
SSA/P	*Sessile Serrated Adenomas/Polyps* - Pólipos/adenomas serrados sésiles
TN	*True Negatives* - Verdaderos negativos
TP	*True Positives* - Verdaderos positivos
TPU	*Tensor Processing Unit* - Unidad de procesamiento tensorial
TSA	*Traditional Serrated Adenomas* - Adenomas serrados tradicionales
WL	*White Light* - Luz blanca

Compendio de publicaciones

Esta tesis es un compendio de las siguientes publicaciones:

[1] Ortega-Morán, J.F.; Azpeitia, Á.; **Sánchez-Peralta, L.F.**; Bote-Curiel, L.; Pagador, J.B.; Cabezón, V.; Saratxaga, C.L.; Sánchez-Margallo, F.M. Medical needs related to the endoscopic technology and colonoscopy for colorectal cancer diagnosis. BMC Cancer (2021) 21, 467.
DOI: 10.1186/s12885-021-08190-z. [IF(2020): 4,430; Q2]

[2] **Sánchez Peralta, L.F.**; Pagador, J.B.; Picón, A.; Calderón, Á.J.; Polo, F.; Andraka, N.; Bilbao, R.; Glover, B.; Saratxaga, C.L.; Sánchez-Margallo, F.M. PICCOLO white-light and narrow-band imaging colonoscopic dataset: a performance comparative of models and datasets. Applied Sciences (2020) 10, 8501.
DOI: 10.3390/app10238501. [IF(2020): 2,679; Q2]

[3] **Sánchez Peralta, L.F.**; Picón A.; Sánchez Margallo F.M.; Pagador J.B. Unravelling the effect of data augmentation transformations in polyp segmentation. International Journal of Computer Assisted Surgery and Radiology (2020) 15, 1975-1988.
DOI: 10.1007/s11548-020-02262-4. [IF(2020): 2,924; Q2]

[4] **Sánchez-Peralta, L.F.**; Picón, A.; Antequera-Barroso, J.A.; Ortega-Morán, J.F.; Sánchez-Margallo, F.M.; Pagador, J.B. Eigenloss: combined PCA-based loss function for polyp segmentation. Mathematics (2020) 8, 1316.
DOI: 10.3390/math8081316. [IF(2020): 2,258; Q1]

[5] **Sánchez-Peralta, L.F.**; Bote-Curiel, L.; Picon, A.; Sánchez-Margallo, F.M.; Pagador, J.B. Deep learning to find colorectal polyps in colonoscopy: a systematic literature review. Artificial Intelligence in Medicine (2020) 108, 101923.
DOI: 10.1016/j.artmed.2020.101923. [IF(2020): 5.326; Q1]

1 Justificación unitaria de la tesis

La detección precoz es actualmente la principal estrategia para reducir la incidencia del cáncer colorrectal y disminuir su impacto en la mortalidad y morbilidad de los pacientes, aumentando considerablemente el porcentaje de supervivencia a los cinco años. En este sentido, las técnicas de aprendizaje profundo (o deep learning) se aplican con gran éxito en el diagnóstico médico y, en concreto, también para mejorar la detección, localización y segmentación de pólipos colorrectales.

Sin embargo, aún existen retos para el uso de aprendizaje profundo en este ámbito, debido en gran medida al tamaño reducido de los conjuntos de datos de imágenes de pólipos colorrectales o la reducida proporción de las lesiones frente a la mucosa sana presente en la imagen endoscópica. Estas situaciones dificultan el correcto entrenamiento de los modelos de aprendizaje profundo. Con el fin de contrarrestar estas limitaciones, puede considerarse la aplicación de técnicas para el aumento de datos o particularizaciones en la función de pérdida, además de una adecuada arquitectura de los modelos.

Por este motivo, en el presente trabajo de tesis se propone analizar en detalle estos problemas para la aplicación de aprendizaje profundo en el ámbito del cáncer colorrectal y proponer soluciones que mejoren su aplicación en detección, localización y segmentación de pólipos colorrectales.

En primer lugar, se ha realizado un análisis de necesidades y una revisión sistemática para posteriormente proponer mejoras en las técnicas de aumento de datos y en la creación de una función de pérdida más adecuada. Además, se ha generado un nuevo conjunto de datos de pólipos colorrectales de acceso abierto que se ha utilizado para comparar resultados de diferentes modelos con otros conjuntos de datos previos también de acceso abierto pero de menor tamaño.

Este trabajo de investigación se ha realizado en el Centro de Cirugía de Mínima Invasión Jesús Usón (Cáceres, España), en estrecha colaboración con el resto de las instituciones que han participado en el proyecto de investigación donde se enmarca la presente tesis, especialmente Tecnalia (Derio/San Sebastián, España). El Centro de Cirugía de Mínima Invasión Jesús Usón es un centro dedicado a la investigación y formación en técnicas quirúrgicas y Tecnalia es Centro de Investigación y Desarrollo Tecnológico. La Unidad de Bioingeniería y Tecnologías Sanitarias del Centro de Cirugía de Mínima Invasión Jesús Usón trabaja actualmente en el desarrollo de algoritmos de aprendizaje profundo con diferentes aplicaciones en el sector sanitario.

2 Introducción

Este capítulo contextualiza los dos temas principales de la presente tesis: el cáncer colorrectal y el aprendizaje profundo. Además, expone los objetivos y presenta brevemente cada una de la contribuciones que forman esta tesis por compendio.

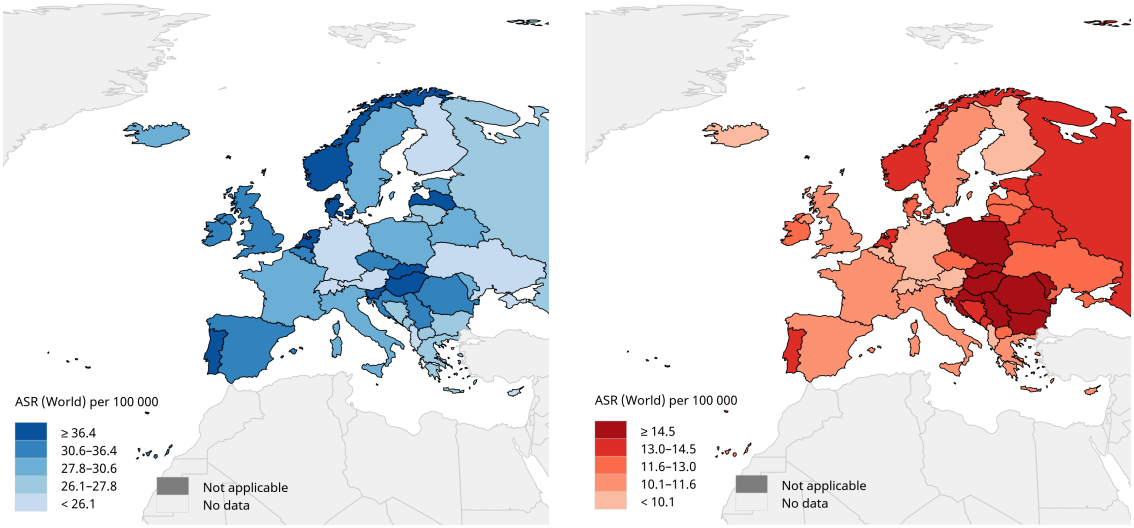
2.1 Cáncer colorrectal

El **cáncer colorrectal** (CCR) se define, de manera general, como un adenocarcinoma localizado en el colon o recto [1]. Se considera una enfermedad relacionada con el estilo de vida, ya que presenta mayor incidencia en países más desarrollados, con dietas altas en calorías y grasas animales así como con una población mayoritariamente sedentaria, y por tanto es altamente prevenible mediante factores modificables con hábitos de vida saludable, entre los que se pueden destacar la realización de actividad física, la reducción del consumo de carne roja y procesada o la reducción de la grasa corporal y abdominal [1, 2]. La mayoría de los casos de CCR se deben a tumores espontáneos originados bien a través de la ruta convencional adenoma-carcinoma o bien a través de la ruta serrada. Estas dos opciones son responsables del 70-75 % y 25-30 % respectivamente de los casos esporádicos de CCR [2]. En menor medida, el CCR puede tener un origen hereditario (5-10 %) o por enfermedad inflamatoria (1-2 %) [3]. Se considera que existen tres alternativas para el desarrollo del CCR [4]:

1. **Secuencia adeno-carcinoma (50 %-70 % del CCR).** La capa mucosa progresivamente evoluciona de un estado normal a una displasia de grado bajo y posteriormente a una displasia de grado alto, para terminar convirtiéndose en neoplasia.
2. **Vía de neoplasia serrada (30 %-35 % del CCR).** Este tipo de lesiones incluye pólipos hiperplásicos (60-75 %), pólipos/adenomas serrados sésiles (SSA/P, de sus siglas en inglés *Sessile Serrated Adenomas/Polyps*) (20-35 %) y adenomas tradicionales serrados (TSA, de sus siglas en inglés *Traditional Serrated Adenomas*) (<1 %). No obstante, todas las lesiones serradas no presentan el mismo potencial para desarrollar CCR. Mientras que los pólipos hiperplásicos tienen menos importancia clínica, los pólipos serrados se consideran lesiones precursoras de CCR [5].
3. **Síndrome de Lynch (3 %-5 % del CCR)**.

En 2020, se estima que a nivel mundial hubo 1,9 millones de nuevos casos de CCR y 935.000 muertes asociadas, representando una décima parte tanto de los nuevos casos de cáncer como de las muertes por esta enfermedad, lo que posiciona al CCR como el tercer tipo de cáncer en términos de incidencia y segundo en términos de mortalidad [6]. Así mismo, se estima que tanto la incidencia como la mortalidad vayan incrementándose anualmente, llegando un millón

de muertes en 2040 [7]. En Europa, se estima que en 2020 se detectaron 519.820 nuevos casos y se produjeron 244.824 muertes (Figura 2.1) [8, 9].



(a) Incidencia. **(b)** Mortalidad.

Figura 2.1: Tasas de incidencia y mortalidad de CCR en Europa [8, 9]. ASR: tasa estandarizada por edad, de sus siglas en inglés *Age-Standardised Rate*.

La sintomatología del CCR incluye no solo síntomas y signos específicos, tales como dolor abdominal, cambios en la frecuencia y textura de las deposiciones o presencia de sangre en heces, sino también otros menos específicos como pueden ser diarrea o estreñimiento, pérdida de peso o cansancio [10]. Afortunamente, la detección temprana del CCR permite aumentar la tasa de supervivencia a 5 años del 18 % cuando se detecta en el grado más avanzado al 88,5 % cuando se detecta en grado inicial debido a los síntomas [11], ya que los tumores detectados en pacientes sintomáticos son mayores y en estadíos más avanzados que los tumores detectados en pacientes asintomáticos en un programa de cribado [12]. Por tanto, mediante estos programas se consigue un aumento significativo de la supervivencia (Figura 2.2), ya que permite detectarlo en niveles menos avanzados y comenzar el tratamiento incluso antes de la aparición de los primeros síntomas, incrementándose la tasa de supervivencia a 5 años del 62,3 % en el caso de detección por síntomas hasta el 83,7 % en el caso de detección por cribado [11].

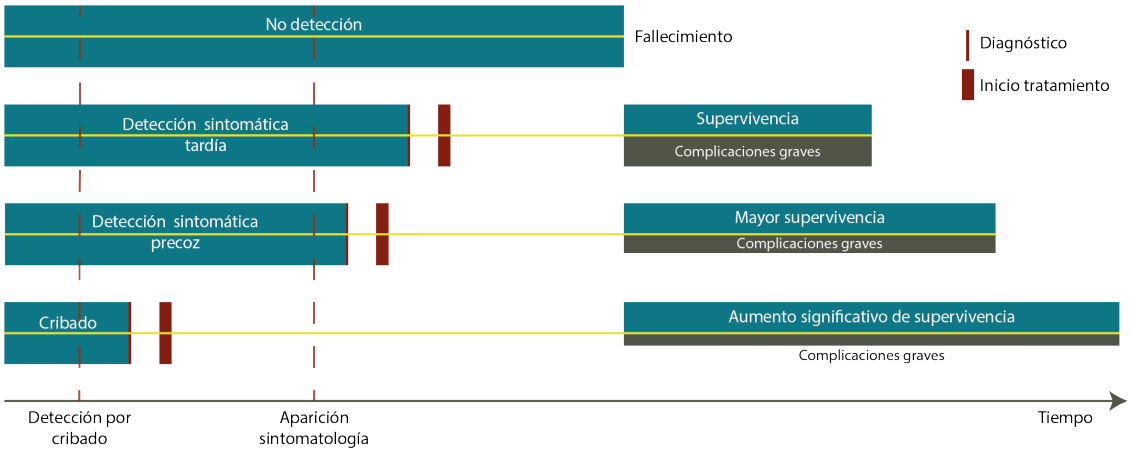


Figura 2.2: Detección precoz sintomática y cribado. Figura basada en [13].

Los **programas de cribado** consisten en la realización de pruebas diagnósticas a personas en principio sanas pero que están dentro de una población considerada diana, con el fin de detectar posibles lesiones precancerosas y minimizar así la incidencia y una mortalidad prematura [14]. La mayoría de los estados miembros de la Unión Europea han implantado programas de cribado en CCR [15, 16] bien a nivel nacional o regional. En España, cada comunidad autónoma establece su propio programa, pero existe bastante consenso en considerar la población diana a aquellas personas entre 50 y 69 años sin antecedentes personales o familiares con criterio de alto riesgo de CCR [17, 18] . Los programas de cribado de CCR suelen consistir en un test de sangre oculta en heces [16]. En caso de dar positivo, se programa una colonoscopia que permita el confirmar o descartar el diagnóstico de CCR y establecer el seguimiento necesario (continuar en el programa de cribado o no, en función de las lesiones identificadas) [17]. La organización de estos programas de cribado aumenta la participación del 38,9 % al 82,7 % a la vez que se reduce la incidencia del CCR un 25,5 % y la mortalidad asociada un 52,4 % [19]. Además, cuanto mayor es el coste de los tratamientos de los casos avanzados de CCR, mayor es el impacto económico del cribado, ya que el coste inicial por paciente diagnosticado en estadio I asciende a 8.644 € mientras que llega a los 13.075 € en estadio III [20].

2.1.1 Colonoscopias y detección temprana de lesiones

La **colonoscopia** es el procedimiento diagnóstico que permite explorar visualmente el colon y recto del paciente mediante la inserción de un endoscopio flexible por el ano [21] y está considerada como técnica estándar para la detección y extirpación de lesiones colorrectales, relacionados con importantes reducciones en la mortalidad por CCR y con complicaciones en un 0,04 % a 8 % de los casos asociados a ciertos factores [22].

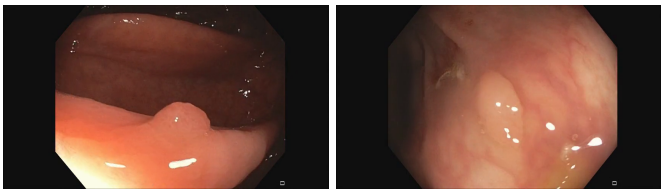
Los pólipos y adenomas (pólipos adenomatosos) son protuberancias de la mucosa colónica en el interior de la luz intestinal y generalmente asintomáticos, y aunque no es frecuente que disminuyan de tamaño, la mayoría presentan un crecimiento mínimo (0.5 mm/año) [23]. No obstante, un aspecto crítico es que una gran parte de las lesiones precancerosas son endoscópicamente imperceptibles [24].

Una vez detectada la lesión, se debe decidir la mejor estrategia a seguir en función del tamaño y forma de la misma. La Sociedad Europea de Endoscopia Gastrointestinal (ESGE, de sus siglas en inglés *European Society of Gastrointestinal Endoscopy*) presentó unas recomendaciones en 2017 que se revisarán en 2022 [25]. Si bien la recomendación general es resecar la lesión y enviarla para un análisis patológico que determine su malignidad, estas recomendaciones introducen las alternativas *"resect and discard"* o "resecar y descartar" y *"diagose and leave behind"* o "diagnosticar y dejar atrás" para pólipos menores o iguales a 5 mm que se diagnostiquen como hiperplásicos ya que presentan bajo riesgo de carcinogénesis, pero siempre que se pueda evaluar su histología en tiempo real con niveles de precisión adecuados.

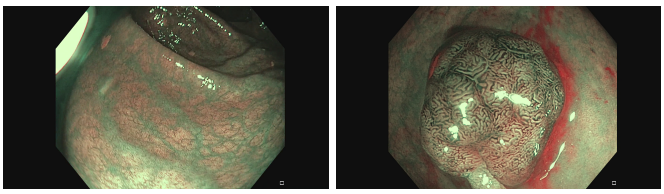
Asociada a la realización de colonoscopias y como indicador de calidad de la misma, la **tasa de detección de adenomas** (ADR, de sus siglas en inglés *Adenoma Detection Rate*) del endoscopista se define como el porcentaje de colonoscopias de cribado realizadas en las que al menos se detecta y reseca un adenoma y cuanto más alta sea, menor es el riesgo de CCR de intervalo [26]. Este se define como el CCR detectado entre los 3 y 36 meses posteriores a una colonoscopia [27] y tiene una incidencia que varía entre el 2,9 y 9,6 % según el estudio, de los cuales, el 58 % podrían explicarse por lesiones no detectadas y un 20 % por una colonoscopia inadecuada [28]. La ADR presenta una relación inversa con el riesgo de CCR de intervalo: una ADR inferior al 20 %

está asociada con un 10 % más de riesgo [29], mientras que un incremento del 1 % en la ADR supone una reducción del 3 % en el riesgo de aparición de CCR [30]. La ADR se ve afectada por factores dependientes del gastroenterólogo, tales como la experiencia o la dedicación exclusiva a la práctica endoscópica [31], y puede fácilmente variar entre el 12,5 % y el 68,1 % [32]. Un endoscopista que realice un volumen medio de colonoscopias y con una ADR entre 7 % y 19 % será responsable de 4,2 CCR de intervalo a lo largo de 35 años de carrera profesional, cifras que dependen mayoritariamente de esos dos factores [33]. Por tanto, es deseable seguir desarrollando sistemas que permitan aumentar la ADR de los endoscopistas, para mejorar la detección temprana y minimizar el riesgo de CCR de intervalo [34].

La colonoscopia con imágenes de luz blanca de alta definición (HD, de sus siglas en inglés *High Definition*) (Figura 2.3a) no ha mostrado mejoras sobre las imágenes de calidad estándard, pero tampoco hay inconvenientes que limiten su uso [35]. De hecho, la ESGE recomienda el uso rutinario de endoscopios HD de luz blanca para detectar neoplasias colorrectales en la población con un riesgo medio, ya que una mínima mejora en la detección tiene un gran efecto debido al alto volumen de procedimientos que se llevan a cabo en dicha población [36]. Otras modalidades de imagen aplican un procesado para mejorar la visualización, como las imágenes de banda estrecha (NBI, de sus siglas en inglés *Narrow Band Imaging*) (Figura 2.3b) que permiten resaltar los patrones vasculares y la superficie de la mucosa [35].



(a) Luz blanca



(b) Imágenes de banda estrecha (NBI)

Figura 2.3: Imagenes extraídas de colonoscopias HD.

No obstante, esta técnica también presenta una serie de limitaciones o condiciones a tener en cuenta, tales como que la calidad de la exploracion presenta gran dependencia con la experiencia del gastroenterólogo que la realiza [31] o que es un procedimiento invasivo para el paciente que puede llegar a causar molestias y que no está exento de riesgos [37]. Además, a la hora de detectar lesiones de apariencia sutil y difícilmente perceptible, hay varios factores a tener en cuenta, entre los que cabe destacar [38]:

- Conocimiento cognitivo y entrenamiento del gastroenterólogo.
    - Identificación de las características visuales de la lesión.
    - Realización de una técnica adecuada.
- Preparación del colon por parte del paciente.
- Utilización de imagénes HD con luz blanca.
- Utilización de tecnologías que faciliten la inspección de la mucosa en zonas ciegas: *third*

eye®retroscope®, endoscopio de amplio espectro (FUSE, de sus siglas en inglés *FUll Spectrum Endoscope*), Endocuff™, EndoRings™, etc.

La ADR además es variable durante el periodo formativo de los gastroenterólogos, sin que la mayoría consiga mejoras tras dicho periodo [39]. Aunque sin embargo, los gastroenterólogos son los únicos facultativos que superan todos los criterios de calidad, incluyendo la ADR, el tiempo de retirada y la tasa de intubación cecal [40].

2.1.2 Tipos de pólipos

De manera general, un pólipo es una protuberancia en la mucosa colónica. Según el tipo de células que la conformen, se pueden identificar [41]:

Pólipos hiperplásicos. Son crecimientos benignos de la mucosa colónica sin potencial de malignidad. Habitualmente son pequeños, inferiores a 5 mm, y localizados en el colon rectosigmoide.

Adenomas colorrectales. Son lesiones premalignas y precursoras del CCR. Pueden presentar displasia de bajo o alto grado y tienen mayor probabilidad de evolución en adenocarcionoma cuanto mayor sea el grado.

Los sistemas de clasificación ayudan a establecer un diagnóstico óptico basado en determinadas características de la lesión, tales como forma, patrones visibles o vascularización. La ESGE recomienda su uso en caso de estar convenientemente validadas [42]. El carcinoma colorrectal avanzado puede dividirse en cuatro tipos, según su apariencia macroscópica (Figura 2.4). A estos subtipos, se añade el carcinoma temprano como tipo 0 [43], ya que es un tumor localizado con potencial para una cura completa tras su resección. Existen diferentes clasificaciones que, si bien ante el uso de diferentes modalidades de imagen pueden no ser intercambiables, sí son complementarias entre sí y pueden usarse de manera combinada según la información que sea necesaria [44].

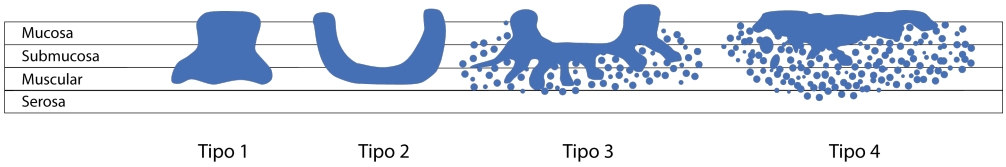


Figura 2.4: Tipos macroscópicos de carcinoma colorrectal avanzado. Figura basada en [43].

La **clasificación de París** se basa en la determinación endoscópica de las lesiones superficiales (no más allá de la capa submucosa) de esófago, estómago y colon [45, 46], permitiendo predecir la profundidad del cáncer superficial y el riesgo de metástasis en ganglios linfáticos [47]. Estas lesiones pueden clasificarse en los siguientes subtipos (Figura 2.5):

- 0-I: polipoide
  - 0-Ip: pedunculada
  - 0-Is: sésil
  - 0-Isp: subpenduculada
- 0-II: no polipoide y no excavada
  - 0-IIa: superficial levemente elevada
  - 0-IIb: plana
  - 0-IIc: superficial levemente deprimida
- 0-III: excavada

Las lesiones 0-III no suelen darse en el colon [48], pero se incluyen en la clasificación de París al aparecer en esófado y estómago. También es posible encontrar combinaciones de estos tipos, como por ejemplo el 0-IIa/c, que es ligeramente elevado por los bordes y con depresión en la zona central. A pesar de ser una clasificación ampliamente extendida, sólo presenta un acuerdo interobservador moderado [49].

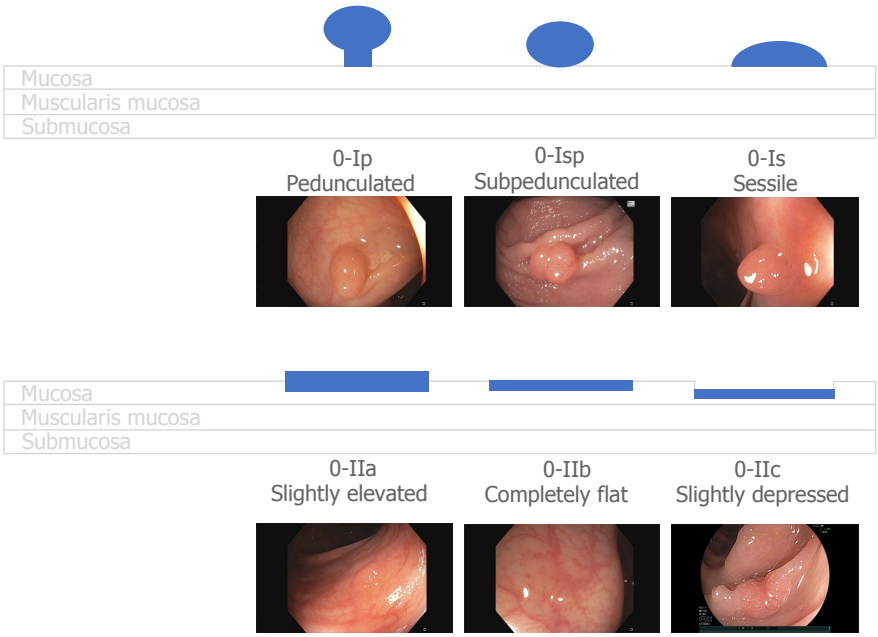


Figura 2.5: Clasificación endoscópica de París para lesiones superficiales [50].

Por último, la **clasificación NICE**, de sus siglas en inglés *NBI International Colorectal Endoscopic Classification* [51], se basa en el color, la vascularización y el patrón de la superfice para clasificar las lesiones en tres tipos (Figura 2.6):

- Tipo I: pólipos hiperplásicos y serrados sésiles.
- Tipo II: adenomas.
- Tipo IIIA y IIIB: cáncer con invasión submucosa profunda.

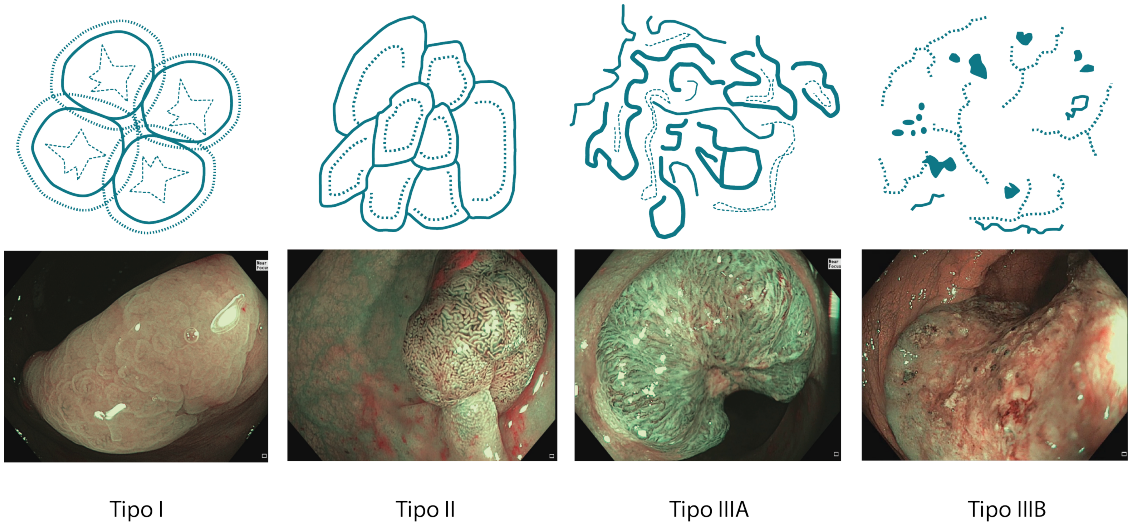


Figura 2.6: Clasificación NICE para pólipos colorrectales. Figura basada en [51].

2.2 Inteligencia artificial y aprendizaje profundo en el procesado de imágenes médicas

Por **inteligencia artificial** (AI, de sus siglas en inglés *Artificial Intelligence*) se entiende un conjunto de tecnologías que permiten que los ordenadores simulen la inteligencia humana en el análisis de datos y es de aplicación habitual en campos como la visión por ordenador, el procesamiento natural del lenguaje o el reconocimiento del habla [52]. Así mismo, dentro del campo de AI, las técnicas de **aprendizaje profundo** (DL, de sus siglas en inglés *Deep Learning*) permiten el aprendizaje de características jerarquizadas de manera automática a partir de los datos en bruto, sin que sea necesario la extracción manual de las características más adecuadas. De este modo, el proceso de aprendizaje es de propósito general (Figura 2.7). Puesto que los niveles de representación se van obteniendo en base a representaciones más simples pero no lineales que van transformando la entrada, es posible la implementación de funciones muy complejas [53]. Esto hace que se extraiga automáticamente la mejor representación interna de los datos, pudiendo incluso explotar información desconocida o sin usar por parte de una persona [54]. En los últimos años, el DL se ha aplicado con gran éxito en multitud de tareas relacionadas con la visión por ordenador, tales como la detección de objetos, reconocimiento de caras o reconocimento de acciones [55]. Este éxito se debe principalmente a los avances tecnológicos en unidades de procesamiento central (CPU, de sus siglas en inglés *Central Processing Units*), unidades de procesamiento gráfico (GPU, de sus siglas en inglés *Graphics Processing Units*) y unidades de procesamiento tensorial (TPU, de sus siglas en inglés *Tensor Processing Units*), la disponibilidad de gran cantidad de datos para entrenamiento y el desarrollo de algoritmos de aprendizaje [56].

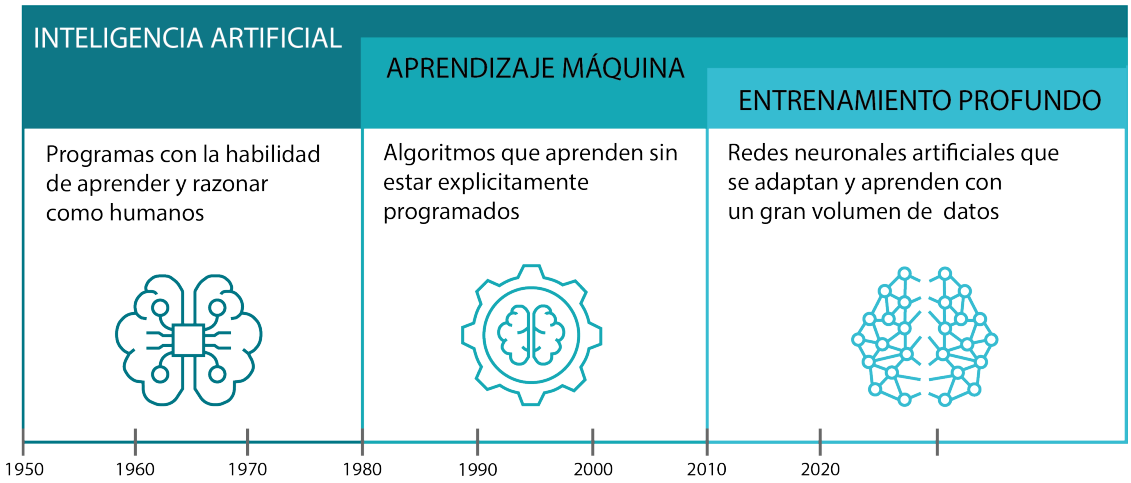


Figura 2.7: Evolución de la inteligencia artificial hacia el aprendizaje profundo.

Como no podía ser de otra manera, el DL también se ha aplicado al análisis de imágenes médicas en multitud de aplicaciones, aunque no exento de grandes retos por la naturaleza única de éstas [57, 58]. Aunque se han llegado a obtener resultados similares e incluso superiores a expertos clínicos [59], sigue siendo necesario llevar a cabo estudios prospectivos y estudios aleatorizados [60], además de considerar cinco principios clave para su adopción en la práctica clínica [61]:

1. El clínico ha de tener el conocimiento mínimo en términos de AI y DL para distinguir hechos de ficción.

2. Las métricas de evaluación del modelo han de ser relevantes en la práctica clínica.
3. Evaluar el posible sesgo del modelo si ha sido entrenado con casos diferentes a los que se van a aplicar.
4. Conocer los parámetros en los que el modelo opera en condiciones seguras.
5. Asegurar que el equipo clínico no desarrolla dependencia del modelo e ignora que puede fallar.

Byrne *et al.* [62] en 2017 ya anticiparon el potencial en las colonoscopias de los sistemas de detección y diagnóstico asistidos por ordenador (CAD, de sus siglas en inglés *Computer Assisted Detection and Diagnosis*), en tres aspectos principales: facilitar una inspección adecuada de la mucosa, mejorar la detección de lesiones y facilitar la predicción histológica de las mismas. Más recientemente, Ashat *et al.* [63], mediante una revisión sistemática de la literatura y un meta-análisis, han concluido que el uso de IA en colonoscopia aumenta tanto la tasa de detección de adenomas (33,7 % frente a 22,9 %) como la de pólipos (45,6 % frente a 30,6 %), mientras Yamada *et al.* [64] han identificado que el uso de sistemas CAD ayuda a los endoscopistas a detectar más lesiones planas que en los grupos de control. Además, también se ha evaluado que el uso de la IA para apoyo de la estrategia "diagnosticar y dejar atrás" supone una reducción de costes de polipectomías y análisis patológicos innecesarios que puede suponer una reducción de entre el 6,9 % y el 18,9 % del reembolso anual bruto para colonoscopias [65]. Por último, cabe mencionar que Ahmad *et al.* [66] han identificado preguntas pendientes por resolver mediante la aplicacion de AI en colonoscopia, agrupadas en cinco bloques temáticos:

1. Métricas de desempeño, diseño de estudios clínicos y criterios de evaluación: organización adecuada de los ensayos para demostrar la eficacia de los modelos de AI/DL, su impacto en el CCR de intervalo o su eficacia en la caracterización de lesiones.
2. Desarrollos tecnológicos: mejora de los modelos para detectar lesiones más complicadas, como pueden ser los pólipos planos.
3. Adopción e integración en la endoscopia: introducción de los modelos de AI/DL en el flujo de trabajo.
4. Datos (acceso, privacidad y anotación): adquisición y anotación de manera eficiente de datos, con menor demanda de intervención humana, para generar la información clínica que se toma como punto de partida, también conocida como verdad terreno.
5. Aprobación regulatoria: aprobación de uso de estos modelos de AI/DL.

2.3 Objetivos

El **objetivo principal** o general de esta tesis es mejorar la detección, localización y segmentación automática de pólipos colorrectales mediante la aplicación de técnicas de aprendizaje profundo.

Este objetivo general puede desglosarse en los siguientes **objetivos específicos** o secundarios para la detección, localización y segmentación automática de pólipos colorrectales:

1. Identificar las necesidades más relevantes para un sistema CAD desde el punto de vista clínico.
2. Analizar detalladamente los diferentes elementos necesarios para implementar modelos de aprendizaje profundo, tales como conjuntos de datos, estrategias de aumento de datos, arquitecturas y métricas empleadas.
3. Diseñar y generar un conjunto de datos y verdad terreno adecuada.
4. Estudiar el efecto de las estrategias de aumento de datos en este tipo de imágenes.
5. Proponer mejoras en la función de pérdida para este campo de aplicación.
6. Diseñar, adaptar y optimizar los modelos basados en aprendizaje profundo más adecuados.

A partir de estos objetivos, esta tesis por compendio recoge los resultados en cinco artículos publicados en diferentes revistas con índice de impacto. A continuación se resumen brevemente cada uno de ellos:

Medical needs related to the endoscopic technology and colonoscopy for colorectal cancer diagnosis. Mediante entrevistas a gastroenterólogos expertos y la realización de una encuesta para ampliar la muestra consultada, este artículo recoge las necesidades médicas relacionadas con la endoscopia y colonoscopia para el diagnóstico de CCR. Las necesidades se han agrupado en: 1) necesidades clínicas, asociadas con una mejora de la detección y clasificación, el tamaño y localización del pólipos o los márgenes y profundidad de penetración; 2) necesidades para los sistemas CAD, que proporcionen información adicional; y 3) necesidades operacionales/físicas, más relacionadas con la limitación de la calidad de imagen de los endoscopios actuales, la flexibilidad de la punta de los endoscopios o una incorrecta preparación por parte del paciente.

Deep learning to find colorectal polyps in colonoscopy: A systematic literature review. Este artículo presenta una revisión sistemática de la literatura. Tras realizar las búsquedas de manera sistemática en diferentes bases de datos y revisar los 1.332 resultados obtenidos en función de los criterios de inclusión y exclusión, se analizó el texto completo de 35 trabajos. A partir de ellos, se identificaron los conjuntos de datos de pólipos colorrectales públicos, concluyéndose que además de tener un tamaño reducido para aplicaciones de DL, no presentaban suficiente variabilidad clínica. Además, se categorizaron los diferentes métodos de DL para detección, localización y segmentación de pólipos colorrectales y para cada categoría se indicaron sus ventajas y desventajas. Se observó que las redes neuronales convolucionales son la arquitectura más utilizada y que los métodos de extremo a extremo son más comunes que los métodos híbridos. En cuanto a métricas, la exhaustividad es la métrica más utilizada para informar de los resultados de los métodos de detección y localización, mientras que para los métodos de segmentación se

usa en mayor medida el índice Jaccard. Finalmente, se concluyó que es necesario desarrollar un marco de referencia común que incluya tanto el conjunto de datos a utilizar como las métricas a calcular y que permita una comparación justa de los diferentes modelos de DL.

Unravelling the effect of data augmentation transformations in polyp segmentation. En este artículo se estudia la influencia de diferentes transformaciones y rangos en los resultados de segmentación de una arquitectura basada en U-Net, entrenada sobre los conjuntos de datos de CVC-EndoSceneStill y Kvasir-SEG. Se consideraron transformaciones a nivel de imagen (desplazamiento en el alto y ancho, rotación, cizallamiento, aumento, reducción, volteo horizontal y vertical y deformación elástica), transformaciones a nivel de pixel (cambios en el brillo y en el contraste) y transformaciones basadas en problemas de la imagen endoscópica (luces especulares e imágenes borrosas). Las transformaciones basadas en pixel mejoraron los resultados de segmentación en CVC-EndoSceneStill, mientras que para Kvasir-SEG las transformaciones a nivel de imagen fueron más adecuadas. Las transformaciones basadas en problemas se comportaron de manera similar en los dos conjuntos de datos. Además, se identificó que tanto el tamaño del pólipo como el brillo y contraste del conjunto de datos tienen gran influencia en estas diferencias. Por tanto, los resultados obtenidos permiten escoger las transformaciones más adecuadas para optimizar la estrategia de aumento de datos para el entrenamiento de modelos de segmentación de pólipos colorrectales.

Eigenloss: combined PCA-Based loss function for polyp segmentation. La definición de la función de pérdida *eigenloss* se presenta en este artículo. Para ello, se han implementado cuatro modelos diferentes, como combinación de dos arquitecturas (U-Net y LinkNet) y dos redes troncales (VGG-16 y Densenet121). Estos cuatro modelos se han entrenado usando siete funciones de pérdida individuales y se ha hecho uso de un análisis de componentes principales para definir los coeficientes de una funcion de pérdida primaria no redundante. En base a estos coeficientes, la *eigenloss* se define como la combinación lineal de las pérdidas individuales usando los elementos del eigenvector como coeficientes. Los resultados mostraron que la *eigenloss* proporciona mejores resultados que las funciones individuales y que es superior a otras combinaciones lineales cuando se usa Linknet. La *eigenloss*, por tanto, permite proyectar las funciones individuales en una única función estadísticamente independiente.

PICCOLO white-light and narrow-band imaging colonoscopic dataset: a performance comparative of models and datasets. El conjunto de datos de luz blanca y NBI PICCOLO se presenta en este artículo, junto con una comparativa de modelos basados en DL. La metodología para la adquisición y anotación de las imágenes, así como las características clínicas de las lesiones del conjunto de datos se exponen en detalle. El conjunto de datos PICCOLO incluye 3.433 imágenes de 76 lesiones de 40 pacientes, anotadas manualmente por gastroenterólogos expertos, y repartidas en conjuntos de entrenamiento, validación y testeo. Los modelos entrenados con el conjunto de datos PICCOLO presentan mejor capacidad de generalización que los mismos modelos entrenados con CVC-EndoSceneStill y Kvasir-SEG. Así mismo, los pólipos planos de categoría 0-IIb según la clasificación de Paris son los que peores resultados obtienen.

3 Técnicas de aprendizaje profundo aplicado a procesado de imagen

En este capítulo se introducen los conceptos de aprendizaje profundo más relevantes en el contexto de la presente tesis. En primer lugar se presenta una breve introducción al aprendizaje profundo y supervisado. Posteriomente se describen los diferentes tipos de capas y arquitecturas más relevantes para los trabajos posteriores y se contextualizan el aumento de datos y las funciones de pérdida.

3.1 Aprendizaje profundo supervisado

El **aprendizaje profundo**, o *deep learning*, se encuadra dentro de la inteligencia artificial y más concretamente dentro del aprendizaje máquina. Mediante DL es posible extraer y aprender de manera automática las características más relevantes de un conjunto de datos [53]. El comportamiento de los modelos de DL y por ende de las redes neuronales se inspira en el comportamiento del cerebro humano (Figura 3.1), en el que las neuronas se comunican entre ellas mediante conexiones sinápticas y permiten procesar los estímulos de entrada para dar lugar a una respuesta, aún siendo las neuronas artificiales una simplificación del funcionamiento real de las biológicas [67].

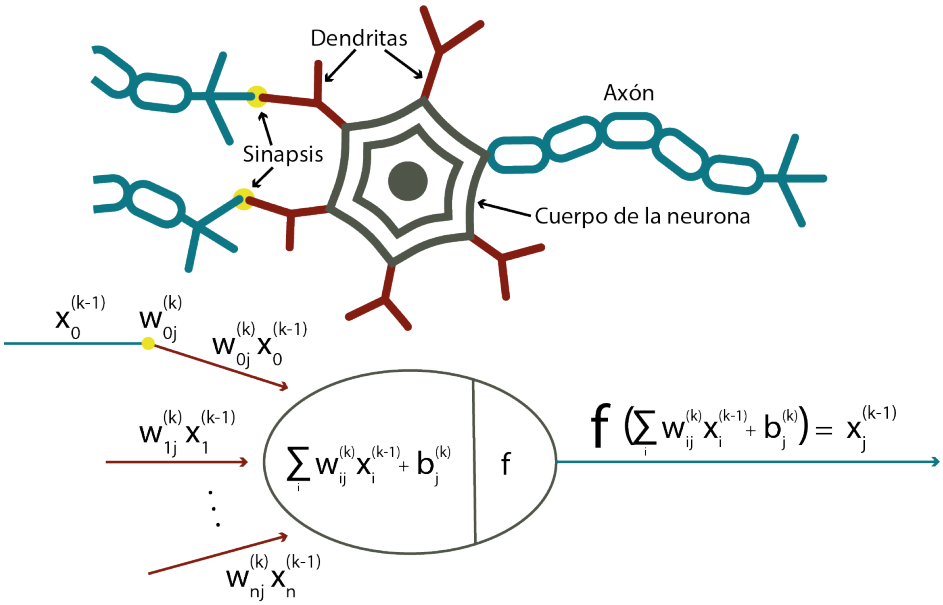


Figura 3.1: Representación esquemática de una neurona biológica y otra artificial.

En la redes neuronales profundas de arquitectura más básica, las neuronas de las diferentes capas están interconectadas entre sí, de tal manera que la neurona j-ésima de la capa k-ésima recibe como entrada el vector de características obtenido en la capa anterior, $k-1$, y sin que haya conexiones entre las neuronas de la misma capa. Los elementos de este vector se combinan mediante una suma ponderada (combinación lineal) según los pesos de la neurona, que indican también la significancia de cada elemento de entrada en dicha neurona. Esta suma es a su vez la entrada a la función de activación, que incluye la no-linearidad y que determina si la neurona se activa o no a la hora de obtener el resultado final, $x_j^{(k)}$. Por tanto, de manera general, la ecuación que rige este comportamiento en una neurona artificial es:

$$x_j^{(k)} = f\left(\sum_i w_{ij}^{(k)} \cdot x_i^{(k-1)} + b_j^{(k)}\right) \tag{3.1}$$

con

$$j \in 1, 2, \ldots, J_k$$
$$k \in 1, 2, \ldots, K$$

y donde

$x_j^{(k)}$ es el valor de salida para la neurona j-ésima de la capa k-ésima,

f es la función de activación,

$w_{ij}^{(k)}$ son los pesos de la neurona j-ésima de la capa k-ésima,

$x_i^{(k-1)}$ son los valores de salida de la capa previa $k-1$,

$b_j^{(k)}$ son los sesgos de la neurona j-ésima de la capa k-ésima,

J_k es el número total de neuronas de la capa k-ésima, y

K es el número total de capas.

Cuando los modelos de DL se aplican sobre una imagen, la información de la capa de entrada a la red, $x = x_j^{(0)}$, son los propios píxeles que conforman la imagen, mientras que la salida de la última capa, $x_j^{(K)}$, se corresponderá con el vector de predicción asociado a la imagen de entrada, $\hat{y}$.

En un esquema de aprendizaje supervisado [53] (Figura 3.2), para cada imagen $\mathbf{X}$ se dispone de una verdad terreno $\mathbf{Y}$, en forma de etiqueta o máscara binaria. Los modelos de DL se someten a entrenamiento iterativo para conseguir mejorar las predicciones, $\hat{\mathbf{Y}}$, y para ello se hace uso de una función objetivo o función de pérdida, $L(\mathbf{Y}, \hat{\mathbf{Y}})$. En cada iteración, esta función calcula el error entre la predicción y el resultado deseado en el conjunto de muestras de entrenamiento en lo que se conoce como *propagación hacia adelante*, ya que la información fluye desde la capa de entrada hasta la capa de salida para poder calcular la predicción. Según el valor obtenido en esta función de pérdidas, los pesos de la red se actualizan mediante un algoritmo de *propagación hacia atrás*, ya que el error se transmite desde la capa de salida a la capa de entrada. Este algoritmo calcula el gradiente negativo de la función de pérdida con respecto a los pesos de la red, identificando la orientación a seguir para llegar al mínimo de la función. Existen diferentes alternativas matemáticas para calcular este gradiente, entre las que caben destacar el descenso de gradiente estocástico (SGD, de sus siglas en inglés Stochastic Gradient Descent), Adam o RMSprop [68] por ser ampliamente utilizadas. Según el valor obtenido en este gradiente, los pesos se modifican levemente (Ecuación 3.2)para reducir el error en la siguiente iteración, t, hasta conseguir llegar a un mínimo.

$$
\begin{aligned}
w_{i,j}^{(k,t)} &= w_{i,j}^{(k,t-1)} - \lambda \cdot \frac{\partial L}{\partial w_{i,j}^{(k,t)}} \\
b_j^{(k,t)} &= b_j^{(k,t-1)} - \lambda \cdot \frac{\partial L}{\partial b_j^{(k,t)}}
\end{aligned}
\tag{3.2}
$$

El tamaño del paso, λ, es lo que se conoce como tasa de aprendizaje. La correcta selección de la tasa de aprendizaje tiene un papel muy relevante en el proceso de entrenamiento, ya que si es muy elevada puede dar lugar a que el modelo no converja, mientras que si es muy baja puede relentizar en exceso la identificación del mínimo de la función de pérdida. Por limitaciones de almacenamiento, dentro de una misma iteración los datos de entrenamiento se pasan a la red en lotes.

A la hora de entrenar estos modelos, también pueden presentarse una serie de problemas relacionados con este proceso de optimización [69]. El sobreajuste, u *overfitting*, se produce cuando el modelo reconoce las características de las imágenes con las que se ha llevado a cabo el entrenamiento con mucha más exactitud que para otras imágenes, lo cual suele suceder debido al tamaño reducido del conjunto de entrenamiento. Otro problema que puede aparecer es el desvanecimiento del gradiente en redes más profundas. En estas redes, es posible que la pérdida decaiga rápidamente a cero en la propagación hacia atrás al no activarse las funciones de activación. Este problema se resuelve en gran medida con la inclusión en la arquitectura de conexiones entre las capas intermedias. Además, la correcta inicialización de los pesos, especialmente en aquellas redes entrenadas desde cero, juega un papel importante para reducir el problema del desvanecimiento del gradiente.

Una vez que el modelo de DL está optimizado y se ha llegado a un mínimo en la función de pérdida, puede emplearse para hacer predicciones sobre conjuntos de datos nuevos para los que se desconoce la verdad terreno.

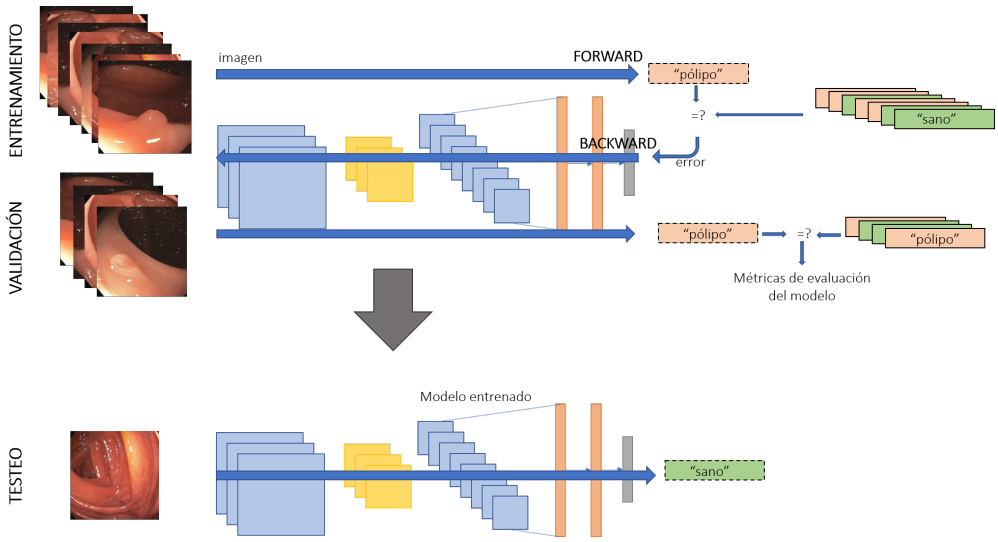


Figura 3.2: Aprendizaje supervisado.

3.2 Arquitecturas

Las arquitecturas de redes de DL están compuestas, por tanto, por múltiples capas no lineales que son capaces de aprender diferentes niveles jerárquicos de abstracción mediante un proceso de optimización iterativo en el que se ajustan las características internas de la red [53]. Los diferentes tipos de capa así como las conexiones que se establezcan entre ellas dan lugar a las diferentes arquitecturas.

3.2.1 Tipos de capas

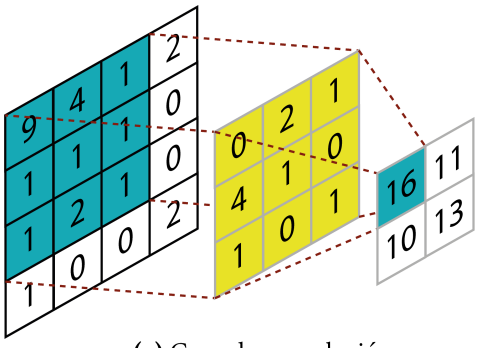
De manera previa a la descripción de diferentes arquitecturas, se describen los principales tipos de capas que pueden emplearse [68, 70–73].

Capa convolucional (CONV). Esta capa (Figura 3.3a) es el tipo básico en cualquier red convolucional. Las capas convolucionales aplican la convolución de un filtro o *kernel* a los datos de entrada. El resultado es un mapa de activación, que tiene como valor el resultado de la multiplicación de los valores de la matriz de entrada por los del filtro (que son parámetros entrenables de la red) según se va desplazando por toda la matriz. En cada capa se aplicará un número determinado de filtros con pesos compartidos, lo que reduce los parámetros entrenables totales. Cada filtro se activará ante una característica diferente de la entrada y por tanto darán lugar a diferentes mapas de activación. Las capas convolucionales permiten aprender relaciones espaciales entre los píxeles de la imagen.

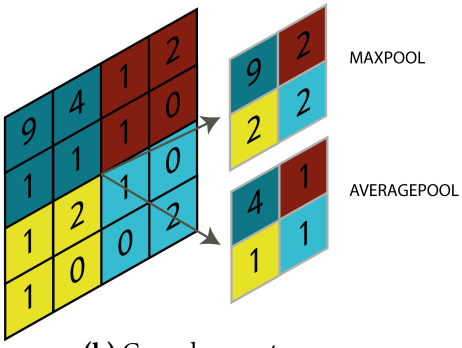
Capa de muestreo (POOL). Esta capa (Figura 3.3b) tiene como objetivo reducir la dimensión del mapa de activación que recibe como entrada, para así reducir el número de parámetros entrenables de la red. Habitualmente se emplea un muestreo del valor máximo (MAXPOOL), en el que se elige el valor más alto de la ventana como valor representativo, frente a otros tipos de muestreo como la suma o la media. Esta capa reduce las dimensiones pero no la profundidad del mapa de activación. Al reducir las dimensiones de los mapas de activación, se refuerza la independencia del modelo a pequeñas variaciones espaciales al mismo tiempo que se comprime la información del mapa de activación. Estas capas se suelen colocar tras una capa convolucional.

Capa de convolución transpuesta. Esta capa (Figura 3.3c) realiza un sobremuestreo, incrementando las dimensiones del mapa de activación a la vez que reduce su profundidad. Habitualmente suele llamarse capa de deconvolución (DECONV) debido a que la entrada es de menor dimension que la salida, al contrario que en la capa CONV, que reduce las dimensiones del mapa de activación. No obstante, es importante puntualizar que la capa DECONV que no es la operación matemática inversa a la convolución.

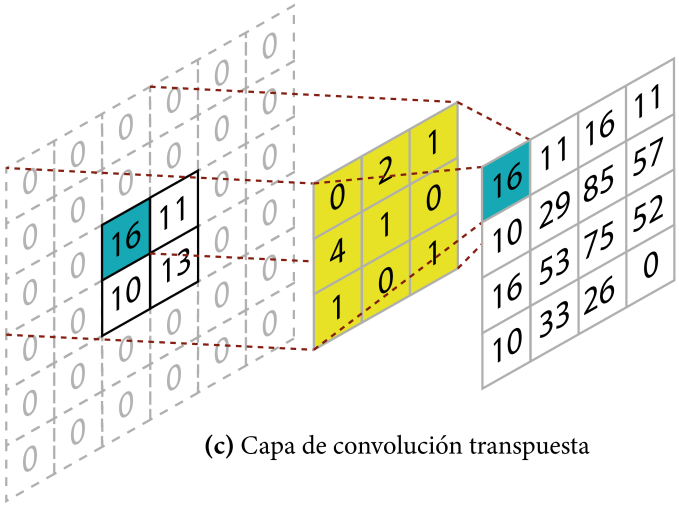
Capa de no linearidad o capa de activación. La función de activación mencionada anteriormente puede considerarse también una capa (Figura 3.3d), responsable de introducir la no linearidad en el modelo de DL que posibilita el aprendizaje de tareas complejas. Si bien puede ser cualquier función no lineal diferenciable (necesario para el cálculo del gradiente y la propogación del error explicado anteriormente), las funciones de activación más utilizadas son la unidad lineal rectificada (ReLU, de sus siglas en inglés *Rectified Linear Unit*), tanh y sigmoidea. Estas funciones activan o desactivan la neurona, dando por tanto énfasis a aquellas características más relevantes para la mejora de la predicción.



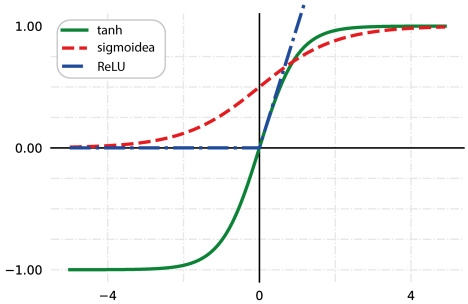
(a) Capa de convolución



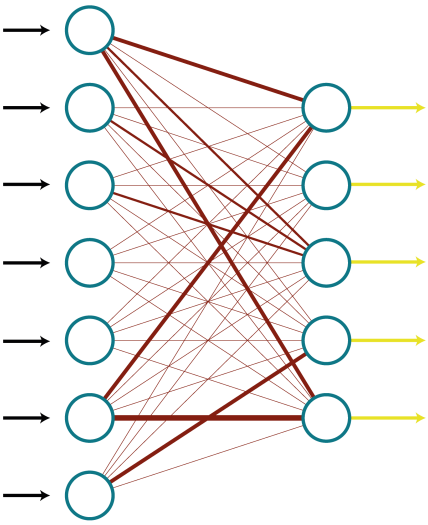
(b) Capa de muestreo



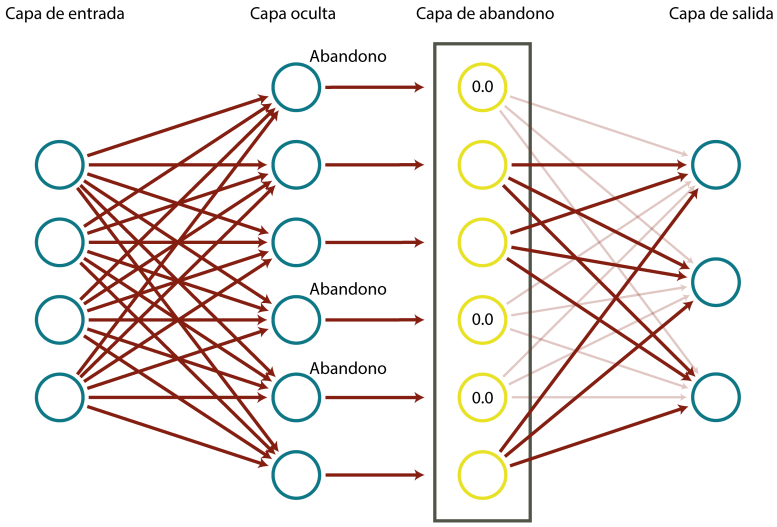
(c) Capa de convolución transpuesta



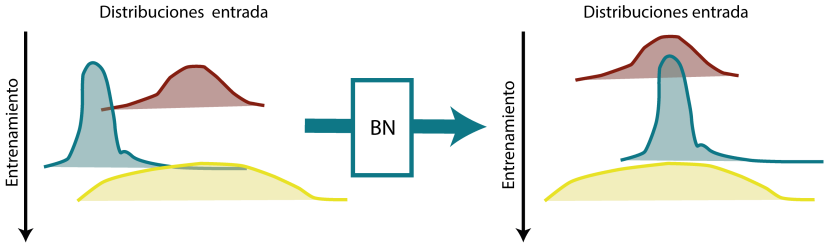
(d) Capa de activación



(e) Capa completamente conectada



(f) Capa de abandono



(g) Capa de normalización por lotes

Figura 3.3: Representación esquemática de los diferentes tipos de capas.

Capa completamente conectada (FC). Es la capa básica de las redes neuronales tradicionales (Figura 3.3e), en las que todas las neuronas de la capa k-esima están conectadas a todas las neuronas de la capa anterior. Pueden verse como capas convolucionales con un filtro de dimensiones 1×1. Suelen encontrarse al final de las arquitecturas orientadas a clasificación, ya que se hace corresponder su salida con las diferentes clases y mediante una función de activación exponencial normalizada (SoftMax) se obtienen las probabilidades de que la entrada pertenezca a cada una de las clases.

Capa de abandono (DROPOUT). Esta capa (Figura 3.3f) aporta normalización a la red, mediante el descarte aleatorio de un determinado número de neuronas en cada iteración del entrenamiento. De este modo se obtienen modelos más robustos y con menos problemas de sobreajuste.

Capa de normalización por lotes (BN). Al igual que la anterior, aporta regularización mediante la normalización de las imágenes del lote para que el conjunto de activaciones de salida tenga media igual a cero y varianza unitaria (Figura 3.3g).

3.2.2 Redes neuronales convolucionales

Las **redes neuronales convolucionales** (CNN, de sus siglas en inglés *Convolutional Neural Network*) son arquitecturas de DL diseñadas para procesar datos matriciales, como es el caso de las imágenes, ya que son tres matrices 2D con los píxeles de cada canal rojo, verde y azul (RGB, de sus siglas en inglés *Red, Green, Blue*), basándose en cuatro conceptos clave: conexiones locales, pesos compartidos, muestreo y uso de múltiples capas [53].

La arquitectura típica de una CNN (Figura 3.4) cuenta con una primera parte compuesta de capas CONV y capas POOL. Las capas CONV detectan características locales en diferentes posiciones del mapa de características de entrada, proporcionando invarianza a la traslación, usándose el mismo filtro con los mismos pesos sobre todo el mapa de características. Por otro lado, las capas POOL agrupan características similares en una, reduciendo el tamaño del mapa de características de entrada. Entre una y otra capa se incluye la no-linealidad, conformando un bloque. Tras varios de estos bloques de capas apilados (normalmente dos o tres), se añaden las capas FC, que darán como resultado la probabilidad de que la imagen de entrada pertenezca a cada una de las clases establecidas.

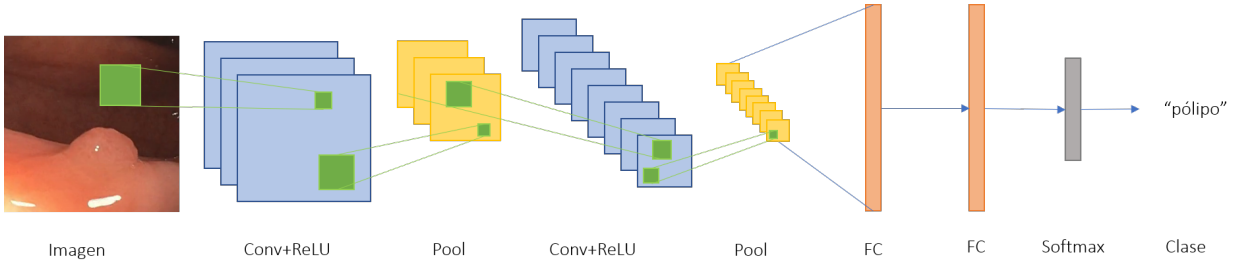


Figura 3.4: Arquitectura básica de una red neuronal convolucional (CNN). Cada tipo de capa se representa con un color diferente: capa convolucional (conv), capa de activación (ReLU), capa de muestreo (pool) y capa completamente conectada (FC).

La arquitectura AlexNet de Krizhevsky *et al.* [74] se convirtió en la primera red CNN que se popularizó al resultar ganadora del *ImageNet Large Scale Visual Recognition Competition* (ILSVRC) en 2012. Esta red presenta ocho capas con parámetros, de las cuales las cinco primeras son capas

CONV y las tres últimas son capas FC, incluyendo capas MAXPOOL y DROPOUT intercaladas en determinadas posiciones. La última de las capas FC da como salida las probabilidades de que la imagen pertenezca a cada una de las 1.000 clases definidas en el conjunto de datos ImageNet. Esta red presenta 60 millones de parámetros entrenables y 650.000 neuronas.

Dos años más tarde, las arquitecturas VGG en sus configuraciones VGG16 y VGG19 propuestas por Simonyan *et al.* [75] consiguieron el segundo puesto en el ILSVRC'14. Aún no obteniendo el mejor resultado, se popularizaron rápidamente al ser modelos simples que se basan en bloques compuestos de capas CONV con filtros de convolución pequeños, de tamaño 3×3, separados por capas MAXPOOL. La red se completa con capas FC y DROPOUT. El uso de estos filtros pequeños permite acoplar más capas, consiguiendo mayor profundidad en la red y por tanto mejores resultados. Por el contrario, la red presenta un gran número de parámetros (138 millones de parámetros para VGG16) y lenta convergencia.

Posteriomente, He *et al.* [76] con una red residual, o ResNet, ganó el ILSVRC'15. Este tipo de redes incluye conexiones de salto (conocidas por su término en inglés *skip connections*) que hacen que la salida de cada bloque residual no sea únicamente la entrada transformada sino que se le suma la entrada original, solventando el problema de degradación asociado a redes más profundas. Similar a las redes VGG, existen diferentes configuraciones según el número de capas, tales como ResNet50 o ResNet101, con diferente número de parámetros entrenables (más de 23 millones para ResNet50).

Más recientemente se han seguido proponiendo diferentes arquitecturas CCN para clasificación de imágenes, como Densenet [77] que basándose en las conexiones de salto de ResNet conecta todas las capas entre sí dentro del mismo bloque denso, por lo que cada capa toma todos los mapas de activación precedentes como entrada, haciéndo la red más robusta al desvanecimiento e incrementado su profundidad. Al igual que las arquitecturas anteriores, se pueden encontrar diferentes configuraciones de Densenet, con 121, 169, 201 o 264 capas, que corresponden con entre 10 y 30 millones de parámetros.

Es frecuente que alguna de estas arquitecturas sea la red troncal de otras arquitecturas más complejas, ya que se aprovechan estos modelos pre-entrenados en grandes conjuntos de datos como es ImageNet para la inicialización de pesos.

3.2.3 Redes neuronales completamente convolucionales

Las CNN requieren que la matriz de entrada se ajuste a un tamaño determinado, establecido por las dimensiones de las capas FC. Para proporcionar mayor flexibilidad, se pueden transformar estas CNN en **redes completamente convolucionales** (FCN, de sus siglas en inglés *Fully Convolutional Networks*) mediante la sustitución de las capas FC por capas CONV (Figura 3.5) y la inclusión de un sobremuestreo mediante una capa DECONV [78]. De esta manera se obtienen redes adecuadas para la segmentación semántica de la imagen de entrada, esto es, el etiquetado a nivel de píxel. Según las características del final de la red, podemos distinguir diferentes arquitecturas: FCN-8s, FCN-16s o FCN-32s, con alrededor de 134 millones de parámetros para la FCN-8s.

Tras el éxito de las FCN para segmentación semántica, comenzaron a surgir arquitecturas codificador-decodificador. En estas arquitecturas, la primera parte de la red, o codificador, transforma la imagen de entrada dando como resultado un vector de características representativo de baja resolución. Por otro lado, la segunda parte de la red o decodificador es responsable de mapear dicho vector en un mapa de igual resolución que la imagen de entrada que determina la

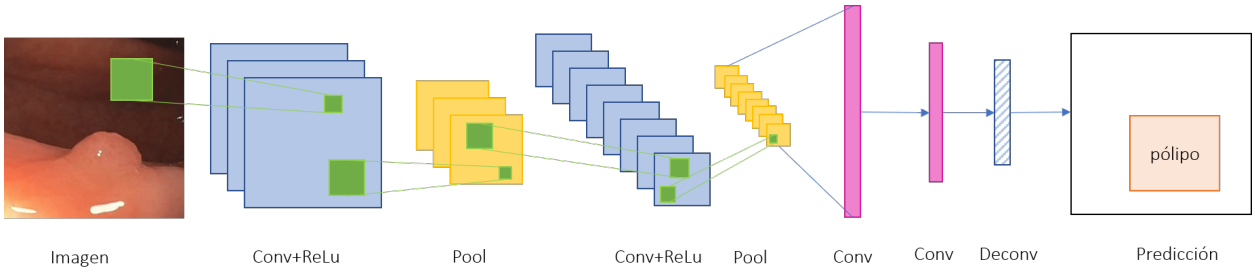


Figura 3.5: Arquitectura básica de una red completamente convolucional (FCN). Cada tipo de capa se representa con un color diferente: capa convolucional (conv), capa de activación (ReLU), capa de muestreo (pool) y capa de convolución traspuesta (deconv).

categorización de cada píxel.

Entre las arquitecturas codificador-decodificador cabe destacar la U-Net propuesta por Ronneberger *et al.* [79]. Diseñada para imágenes biomédicas, la U-Net cuenta con un codificador formado por cuatro módulos convolucionales seguidos de sus correspondientes capas MAXPOOL para reducir las dimensiones. Así mismo, el decodificador presenta otros cuatro módulos convoluciones seguidos por una capa DECONV que permite recuperar el tamaño original. Los bloques equivalentes de codificador y decodificador están conectados mediante conexiones de salto que permiten combinar características de alto y bajo nivel. Para ello, se concatenan los mapas de características del codificador en su nivel correspondiente en el decodificador. A partir de modificaciones sobre la arquitectura básica de la U-Net se han propuesto otras como 3D U-Net, V-Net o Res-UNet [80].

Otras arquitecturas como LinkNet [81] tienen una estructura similar, diferenciándose en la manera de unir la información procedente del codificador, ya que en lugar de la concatenación se suman los mapas de características.

3.3 Aumento de datos

El **aumento de datos**, o *data augmentation*, es una estrategia que permite incrementar de manera artificial el conjunto de datos de entrenamiento, bien porque éste sea reducido o porque las clases se encuentren desequilibradas. El objetivo es minimizar el efecto de memorización de dicho conjunto, mejorando la generalización del modelo [82]. Una de las posibilidades es crear nuevos datos sintéticos, aplicando modificaciones al conjunto de entrenamiento original que se toma como base. Estas modificaciones dependen de la naturaleza de los datos de entrada, ya que los datos sintéticos deben de mantener una similitud con los datos originales. Además, son también dependientes de la tarea que vaya a desempeñar la red, ya que estas aprenden invarianzas en el dominio de los datos. En el caso de imágenes, el aumento de datos se lleva a cabo fácilmente mediante transformaciones tales como rotación, desplazamiento, reflexión o deformación elástica, que resultan en imágenes para entrenamiento similares a un ejemplo concreto, pero no es trivial ante datos como espectrogramas [83] o electroencefalografía [84].

Para la segmentación de imágenes médicas, Tajbakhsh *et al.* [85] establece diferentes tipos de aumento de datos:

- Aumento tradicional, y que según la característica de imagen que se modifique se puede a su vez dividir en:
    - Calidad de la imagen: nitidez, borrosidad y ruido.

    - Apariencia de la imagen: brillo, saturación y contraste.
    - Composición de la imagen: transformaciones espaciales como rotación, escalado y deformación.
- Aumento mixto, en el que las nuevas imágenes y verdad terreno se obtienen como combinación de pares de imágenes del conjunto de entrenamiento.
- Aumento sintético, principalmente basado en redes generativas antagónicas (GAN, de sus siglas en inglés *Generative Adversarial Networks*), y que a grandes rasgos puede dividirse en:
    - Mismo dominio, si se sintetizan imágenes en el mismo dominio.
    - Dominio cruzado, si se proyectan imágenes etiquetadas en otro dominio al dominio objetivo.

El efecto del aumento de datos es tanto mayor cuanto menor es el tamaño del conjunto de entrenamiento [86], pero la elección de las transformaciones y de sus rangos suele hacerse de manera empírica, según la imaginación, tiempo y experiencia del investigador [87]. Además, es preciso tener en cuenta que incrementar en exceso las transformaciones puede ser contraproducente [88].

3.4 Función de pérdida

La **función de pérdida**, o *loss function*, se utiliza comúnmente en problemas de optimización, en los que se busca minimizar la desviación entre los valores estimados y los reales para un conjunto de datos determinado. Por lo tanto, dicha función de pérdida dirige el entrenamiento de los modelos de DL mediante el cálculo del error entre los valores predichos y la verdad terreno y minimizándolo para optimizar los pesos del modelo. La relevancia e influencia de las funciones de pérdida en los modelos de DL se han estudiado en muchos campos, como el reconocimiento de dígitos escritos a mano [89], clasificación única rápida [90], reconstrucción de resonancia magnética cardíaca [91], reducción de ruido en imágenes de TAC de baja dosis [92], o reconocimiento facial [93].

Las funciones de pérdida pueden clasificarse según el aspecto en el que basen la minimización: la distribución, la región o el contorno. En cuanto a las funciones de pérdida basadas en la distribución, es común emplear la entropía cruzada, o *cross entropy* [94]. Esta función calcula el coste en cada píxel usando una función logarítmica que luego se promedia sobre todos los píxeles para cuantificar la diferencia entre las dos distribuciones de probabilidad (verdad terreno frente a predicción). Otras funciones basadas en distribuciones son la entropía cruzada ponderada [79] y la pérdida focal [95]. En el caso de funciones de pérdida basadas en la región, es común utilizar la función Dice [96], ya que mide la superposición de dos conjuntos, lo que puede aplicarse fácilmente a la segmentación binaria. Otras funciones basadas en regiones son la función basada en el índice Jaccard o intersección sobre la unión (IoU, de sus siglas en inglés *Intersection over Union*) [97], la función de pérdidas Tversky [98], de Lovász-Hinge o Lovász-Softmax [99], o la función de pérdidas de Tversky focal [100]. Por último, las funciones basadas en contornos incluyen por ejemplo las funciones de superficie, que emplean una métrica de distancia para comparar las formas de los dos contornos [101], o la pérdida que utiliza las estimaciones de la distancia de Hausdorff para medir las desviaciones entre dos contornos [102].

Cuando el DL se aplica al análisis de imágenes médicas, se suelen utilizar las funciones de pérdida de entropía cruzada junto con una segunda medida de distancia o superposición [103],

por lo que se utilizan comúnmente las funciones de pérdida combinada [104–106]. Para la segmentación de pólipos, se pueden encontrar tanto funciones de pérdida individuales como combinadas. Wichakam *et al.* [107] aplican la función de pérdidas de Dice sola. Por otra parte, mientras que Zhou *et al.* [108] combinan la función de pérdidas de Dice con la entropía cruzada binaria, Mohammed *et al.* [109] la combinan con la entropía cruzada binaria ponderada en su lugar.

4 Detección, localización y segmentación automática de pólipos colorrectales basadas en técnicas de aprendizaje profundo

Este capítulo describe las tareas objetivo, así como contextualiza los trabajos realizados en función del estado del arte actual y sus limitaciones. Además, se presentan los conjuntos de datos y las métricas empleadas en los diferentes trabajos.

4.1 Métodos de detección, localización y segmentación

En primer lugar, es necesario definir las tareas que se pueden llevar a cabo en el contexto de sistemas CAD para pólipos colorrectales (Figura 4.1):

Detección. Identificar si en una imagen o fotograma dado existe un pólipo o no, sin dar información de dónde se encuentra este.

Localización. Determinar la posición del pólipo en la imagen o fotograma, sin que la forma exacta de este sea relevante.

Segmentación Marcar de forma precisa el área de la imagen que se considera pólipo.

Clasificación. Para un pólipo dado, categorizarlo como benigno o maligno o establecer su categoría según algún esquema de clasificación (París, NICE, etc.).

En esta tesis, el foco está puesto en la detección, localización y segmentación de pólipos, quedando la clasificación fuera del alcance de la misma. El nivel de información aportado por estas tareas, así como de la verdad terreno necesaria, va incrementándose. Por tanto, a partir de la segmentación de un pólipo podemos determinar tanto la detección como la localización, no siendo posible la situación inversa, es decir, determinar la localización y segmentación a partir de la detección. Es por esto que los trabajos de esta tesis estarán orientados en gran medida a la segmentación de pólipos. Asímismo, es importante mencionar que los términos imagen y fotograma se usarán indistintamente a lo largo del resto del documento.

Como no podía ser de otra manera, los métodos de DL también tienen gran presencia en el campo de la detección, localicación y segmentación de pólipos colorrectales. Ya en 2015, en el sub-challenge *Automatic Polyp Detection* organizado como parte del *Endoscopic Vision Challenge* [110] en el marco del congreso de relevancia internacional *Medical Image Computing and Computer Assisted Intervention* (MICCAI), se presentaron varias propuestas de métodos de DL. Estos, basados mayoritariamente en CNN, sobrepasaron claramente los resultados de los

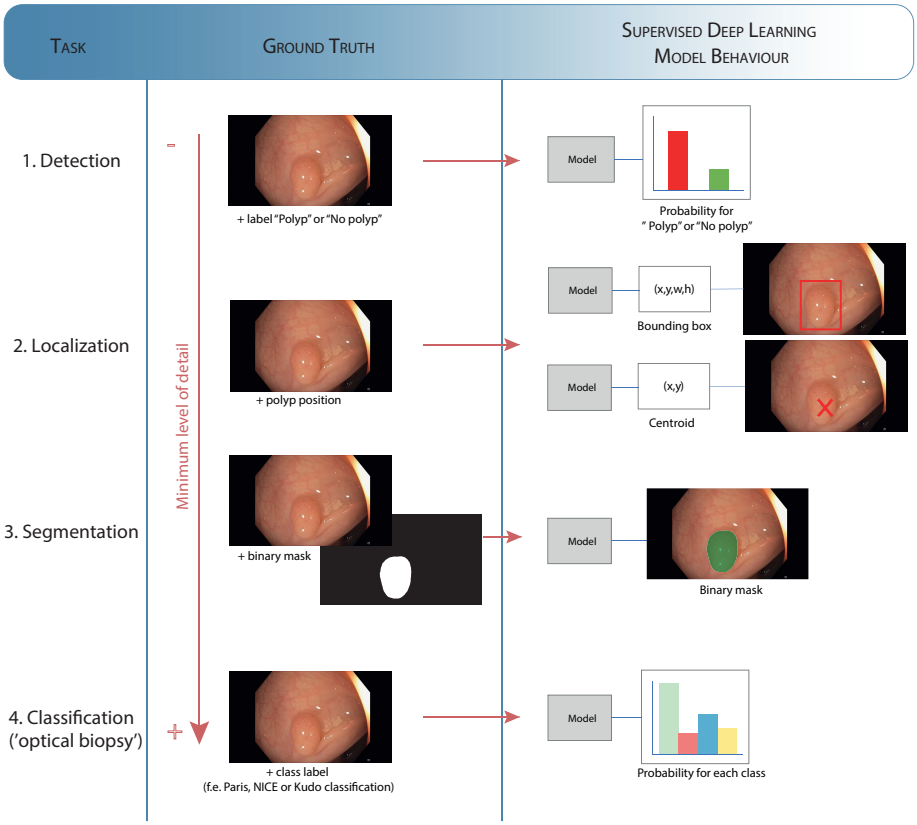


Figura 4.1: Representación esquemática de las tareas consideradas en esta tesis: detección, localización y segmentación de pólipos colorrectales.

métodos tradicionales en los que las características se extraían manualmente, posicionando estas arquitecturas como estado del arte [111].

Desde entonces el número de métodos propuestos ha ido creciendo. En este sentido, la Figura 4.2 muestra los resultados por años obtenidos en una búsqueda en la Web of Science con la cadena de búsqueda

((“colon” OR “colorectal”) AND (“cancer” OR “polyp”) AND (“deep learning”) AND (“detection” OR “localization” OR “segmentation”))

donde se puede ver claramente la evolución exponencial de los trabajos publicados en este campo desde el 2015 hasta la actualidad.

El aumento de trabajos junto con el creciente interés de la comunidad científica para aplicar sistemas basados en AI/DL en la práctica clínica diaria ha dado como resultado la proliferación de revisiones, sistemáticas o no, del estado del arte. Solo en 2020 y 2021 es posible encontrar multitud de revisiones de la literatura con diferentes enfoques. Varias de estas revisiones analizan los métodos de DL en función de su influencia en la mejora de la ADR [63, 112–115], incluso de manera concreta para pólipos planos [64, 116], o los comparan con expertos [117, 118]. Otras revisiones por el contrario se centran en aspectos más técnicos de los modelos pero están enfocadas a usuarios clínicos por lo que se da mayor énfasis a aspectos como el ADR [119–128]. Por último, otras revisiones se centran en aspectos principalmente técnicos de los métodos de DL [129, 130] entre las que además cabe destacar las de Sánchez-Peralta *et al.* [50], Pacal *et al.* [72] y Nogueira-Rodriguez *et al.* [131], que profundizan más en las arquitecturas de los diferentes métodos, así como en otros aspectos técnicos relevantes relacionados: conjuntos de

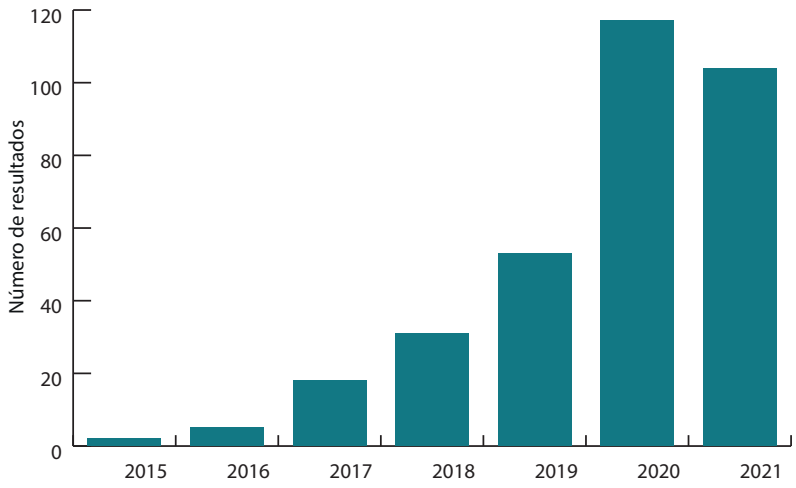


Figura 4.2: Evolución temporal del número de publicaciones sobre detección, localización y segmentación de pólipos colorrectales. Búsqueda realizada el 31 de agosto de 2021.

datos, estrategia de aumento de datos, métricas o limitaciones y problemas que se presentan a la hora de realizar la comparación de métodos. En estas revisiones se pone de manifiesto la gran variedad de arquitecturas y modelos que pueden emplearse para las tres tareas definidas. A partir de ellas y de manera general, puede decirse que los modelos de DL para pólipos colorrectales son mayoritariamente arquitecturas entrenadas de extremo a extremo, en las que la presencia de CNN es significativamente mayor sobre todo para detección y localización, seguidas de las arquitecturas codificador-decodificador para segmentación semántica.

4.2 Limitaciones de los métodos actuales

No obstante y a pesar de la gran variedad de métodos y resultados positivos en muchos de ellos, aún existen ciertas limitaciones identificadas en diferentes revisiones:

1. La mayoría de los trabajos se basan en conjuntos de datos públicos de tamaño limitado o conjuntos de datos privados [50, 131]. Si bien los conjuntos de datos públicos facilitan la comparación de métodos al ser de acceso abierto, suelen tener tamaño limitado y no proporcionan información clínica de las lesiones que incluyen. Por otro lado, el uso de conjuntos privados de datos pueden solventar la limitación del número de imágenes, pero imposibilita la comparación al presentar resultados calculados sobre imágenes propietarias.
2. Aunque es habitual usar estrategias de aumento de datos, no existe una metodología o recomendación sobre cuales son las mejores transformaciones y rangos a aplicar [50, 131]. La elección suele quedar al criterio y experiencia del investigador, sin que haya una base para tal elección.
3. La selección de la función de pérdida no se argumenta, a pesar de la gran influencia de esta en los resultados del modelo de DL, especialmente en las situaciones desequilibradas en la que la clase positiva o de interés es muy inferior a la clase negativa [50].
4. No existe un marco de validación común que se base en un conjunto de datos anotado de grandes dimensiones y acceso público, y que determine también las métricas a calcular, por lo que diferentes autores presentan sus resultados sobre diferentes conjuntos de testeo y haciendo uso de diferentes métricas, imposibilitando una comparación justa de los métodos [50, 72, 123, 131].

4.3 Conjuntos de datos de pólipos colorrectales

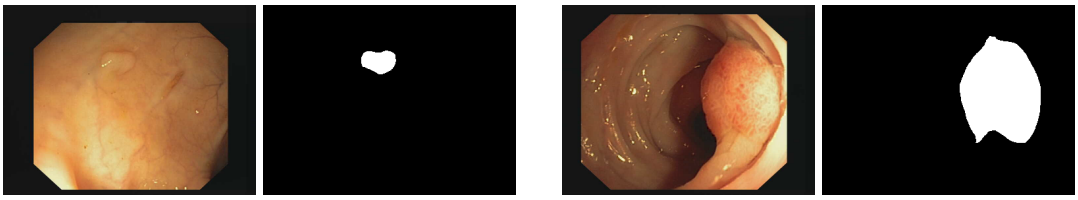
Esta sección presenta los conjuntos de datos de pólipos colorrectales utilizados en esta tesis. A pesar de ser la creación de un conjunto de datos mejorado uno de los objetivos, en las fases iniciales fue necesario utilizar conjuntos de datos de acceso público, aún cuando presentasen ciertas limitaciones.

4.3.1 Conjunto de datos públicos

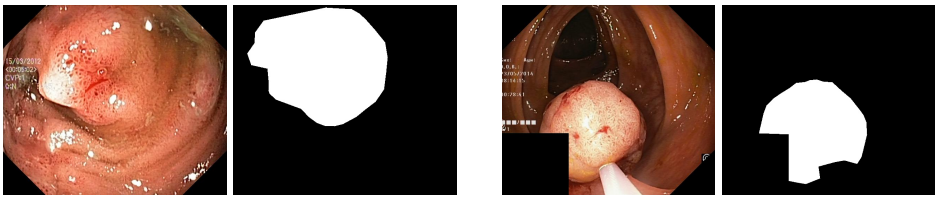
CVC-EndoSceneStill [132] (Figura 4.3a) es un conjunto de datos que proporciona una segmentación manual de las imagenes y que ha sido creado a partir de dos conjuntos de datos anteriores, CVC-ColonDB y CVC-ClinicDB. CVC-EndoSceneStill incluye 912 images de 44 secuencias de video de 36 pacientes. Para cada imagen, se dispone de una máscara binaria para la luz intestinal, las luces especulares, el pólipo y la zona no válida de la imagen. Además, los autores del conjunto de datos propocionan la distribución de las imágenes entre los subconjuntos de entrenamiento, validación y testeo.

Kvasir-SEG [133] (Figura 4.3b) proporciona una máscara binaria manualmente delimitada para 1.000 pólipos, junto con un fichero JSON en el que se determinan los correspondientes cuadros delimitadores. Este conjunto de datos es a su vez parte de otro mayor y de propósito más general para endoscopia gastrointestinal llamado HyperKvasir [134].

En ambos casos, las máscaras binarias muestran con píxeles blancos la región correspondiente a la clase que se indique en la máscara y con píxeles negros se denota el fondo (Figura 4.3). Así mismo, ninguno de los dos conjuntos de datos proporciona metadatos clínicos de las lesiones.



(a) CVC-EndoSceneStill



(b) Kvasir-SEG

Figura 4.3: Ejemplos de las imágenes proporcionadas por CVC-EndoSceneStill y Kvasir-SEG.

4.3.2 Conjunto de datos de PICCOLO

El conjunto de datos de luz blanca y NBI PICCOLO [135] incluye 3.433 imágenes de pólipos obtenidos de 76 lesiones de 40 pacientes. Es actualmente el conjunto de datos de acceso público con mayor número de imágenes manualmente segmentadas por expertos. También se proporciona la distribución de las lesiones en los subconjuntos de entrenamiento, validación y

testeo para asegurar la independencia a nivel de pacientes entre los subconjuntos y facilitar la comparación de los métodos que lo utilicen. Para todas las imágenes, se incluye una máscara binaria manualmente delimitada por gastroenterólogos expertos. Además, es el único conjunto de datos que proporciona metadatos clínicos para cada lesión. Este conjunto de datos se presenta en detalle en la Sección 5.5 al tratarse de una contribución de esta tesis.

4.4 Métricas

Para evaluar los resultados de los modelos de DL es necesario identificar las métricas más adecuadas en función del tipo de tarea que se va a llevar a cabo. Estas métricas comparan las predicciones realizadas por el modelo con la verdad terreno identificada por los expertos. Dicha verdad terreno puede estar dada mediante etiquetas, recuadros delimitadores y/o máscaras binarias. En cualquier caso, las métricas suelen basarse en los elementos de la matriz de confusión (Figura 4.4) [136]. Los cuatro elementos básicos se definen como:

1. Verdaderos positivos (TP, de sus siglas en inglés *True Positives*): número de elementos de tipo pólipo correctamente identificados como pólipo.
2. Verdaderos negativos (TN, de sus siglas en inglés *True Negatives*): número de elementos de tipo no pólipo correctamente identificados como no pólipo.
3. Falsos positivos (FP, de sus siglas en inglés *False Positives*): número de elementos de tipo no pólipo incorrectamente identificados como pólipo.
4. Falsos negativos (FN, de sus siglas en inglés *False Negatives*): número de elementos de tipo pólipo incorrectamente identificados como no pólipo.

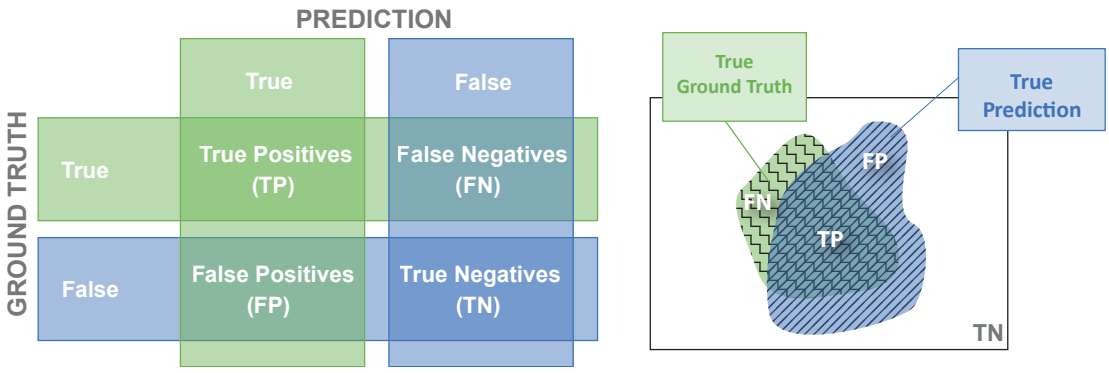


Figura 4.4: Matriz de confusión. A la izquierda, calculada a nivel de imagen. A la derecha, calculada a nivel de pixel por superposición de dos máscaras binarias [50].

Esta matriz de confusión puede calcularse a nivel de imagen o de píxel. Mientras que el primer caso suele hacerse para tareas de detección y localización, el segundo es más habitual en modelos de segmentación. Por tanto, a partir de estos cuatro elementos básicos, se pueden definir las siguentes métricas [136]:

1. Exactitud o *accuracy* (ACC). Proporción de elementos correctamente clasificados frente al total de elementos. De forma general se calcula como

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.1}$$

aunque también es posible encontrarla como

$$Acc = \frac{TP}{TP + TN + FP + FN} \tag{4.2}$$

que puede interpretarse como la exactitud para una clase determinada.

2. Precisión o *precision* (PREC). Proporción de elementos positivos correctamente identificados como tal sobre el total de elementos positivos obtenidos.

$$Prec = \frac{TP}{TP + FP} \tag{4.3}$$

3. Exhaustividad o *recall* (REC). Proporción de elementos positivos correctamente identificados como tal sobre el total de elementos positivos existentes.

$$Rec = \frac{TP}{TP + FN} \tag{4.4}$$

4. Especificidad o *specificity* (SPEC). Proporción de elementos negativos correctamente identificados como tal sobre el conjunto de elementos negativos.

$$Spec = \frac{TN}{TN + FP} \tag{4.5}$$

5. Valor F2 o *F2-score* (F2). Combina la precisión y la exhaustividad en una única métrica.

$$F2 = \frac{5 \cdot Prec \cdot Rec}{4 \cdot Prec + Rec} = \frac{5 \cdot TP}{5 \cdot TP + FN + FP} \tag{4.6}$$

6. Índice Jaccard o intersección sobre la unión (IoU). Indica el porcentaje de solape entre la máscara de la verdad terreno y la máscara de predicción, estando relacionada con el índice Dice directamente.

$$IoU = \frac{TP}{TP + FP + FN} = \frac{Dice}{2 - Dice} \tag{4.7}$$

7. Índice Dice. Indica el grado de similitud entre dos regiones, en este caso, la máscara de la verdad terreno y la máscara de predicción, estando relacionado directamente con el índice Jaccard.

$$Dice = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} = \frac{2 \cdot IoU}{1 + IoU} \tag{4.8}$$

No obstante, varias revisiones [50, 72, 131] ponen de manifiesto la disparidad a la hora de utilizar las métricas por parte de los diferentes autores, lo que hace más complicado una comparación justa de los métodos de DL. En cualquier caso, es recomendable proporcionar más de una métrica como resultado final de los modelos de DL [72] e identificar y seleccionar las métricas de una manera adecuada, ya que una elección incorrecta puede dar lugar a un incremento artificial de los resultados. La Figura 4.5 muestra una situación en una tarea de segmentación donde el conjunto de testeo está compuesto por seis imágenes de tres pólipos diferentes. Mientras que la exactitud obtiene valores aceptables superiores a 0,75 en todos los casos, al incluir los TN en el numerador, el índice Jaccard refleja las diferencias en la segmentación entre las seis imágenes. De manera similar, si las seis imágenes se analizan en conjunto, sin tener en cuenta su origen, se obtendrían valores de exactitud e índice Jaccard medios, pero no podría identificarse que es el pólipo 3 el que obtiene los peores resultados. Si, por otro lado, consideramos la tarea de detección

en base a los resultados de segmentación obtenidos, determinando que un pólipo se detecta cuando el índice Jaccard está por encima del umbral de 0,5, se obtienen resultados diferentes si se analizan las seis imágenes en conjunto o si se analizan por pólipo, decrementándose la capacidad del 50 % de las imágenes al 33 % de los pólipos.

	POLYP 1			POLYP 2	POLYP 3		
	Frame A	Frame B	Frame C	Frame D	Frame E	Frame F	
TP	20	15	17	17	5	5	
TN	80	75	78	58	80	70	
FP	0	5	5	10	0	15	
FN	0	5	0	15	15	10	Average
Accuracy	1.000	0.900	0.950	0.750	0.850	0.750	0.867
per polyp	0.950			0.750	0.800		0.833
Jaccard	1.000	0.600	0.773	0.405	0.250	0.167	0.532
per polyp	0.791			0.405	0.208		0.468
detection (J>0.5)	1	1	1	0	0	0	0.500
per polyp	1			0	0		0.333

Figura 4.5: Comparación de métricas. TP: verdaderos positivos. TN: verdaderos negativos. FP: falsos positivos. FN: falsos negativos. La exactitud se calcula según Ecuación 4.1, mientras que el índice Jaccard se calcula según la Ecuación 4.7. Las líneas punteadas indican la verdad terreno.

5 Contribuciones de esta tesis

En este capítulo se revisan de manera resumida todos los artículos que conforman esta tesis, haciendo especial énfasis en los resultados y contribuciones aportadas. Para más detalles, el texto completo de los artículos se incluye en la Sección 8 y el material suplementario de cada uno de ellos en el Apéndice A.

5.1 Necesidades médicas relacionadas con la tecnología endoscópica y la colonoscopia para diagnóstico de cáncer colorrectal

El primer artículo de esta tesis por compendio recoge un análisis de las necesidades médicas relacionadas con las tecnologías endoscópicas y la colonoscopia. Para ello se llevaron a cabo entrevistas semi-dirigidas con seis endoscopistas expertos y posteriormente se preparó un cuestionario online para recopilar la opinión de un mayor número de endoscopistas.

Tras el análisis de las transcripciones de las entrevistas mediante el método de asignación latente de Dirichlet (LDA, de sus siglas en inglés *latent Dirichlet allocation*), se identificaron como ejes principales las palabras *paciente* y *pólipo*. En relación al segundo, pone de manifiesto la relevancia clínica de una correcta detección e identificación de los pólipos en la imagen endoscópica.

En total se obtuvieron 103 cuestionarios completos, con representación de endoscopistas principalmente de Europa (95,2 %), pero con un pequeña representación también de América del Sur, Asia y África. Los resultados del cuestionario relacionados con el sistema CAD se recojen en la Tabla 5.1.

En base a los resultados de las entrevistas y cuestionarios, se identificaron tres categorías de necesidades médicas: 1) necesidades clínicas, 2) necesidades del sistema CAD y 3) necesidades operacionales/físicas. Aunque se recoge una amplia variedad de necesidades, los aspectos más relevantes relacionados con la detección, localización y segmentación de pólipos que se detectaron son, para cada una de las categorías, los siguientes:

1. **Necesidades clínicas.** La principal preocupación de los endoscopistas es mejorar la detección de los pólipos, especialmente los planos, para así reducir la tasa de pólipos perdidos. Demandan también herramientas que les ayuden a mejorar la detección, considerándose estas especialmente útiles para los pólipos planos mencionados anteriormente, ya que son más difíciles de detectar que los sésiles o pedunculados.
2. **Necesidades del sistema CAD.** En este sentido, los endoscopistas requieren que los sistemas CAD les aporten información adicional para mejorar la detección y diagnóstico de los pólipos. Esta información ha de proporcionarse de manera visual y preferentemente

Tabla 5.1: Necesidades clínicas y del sistema CAD. Datos extraídos de las tablas 1 y 2 de [137].

	N	%
Sería deseable tener un sistema automático	81	88,0
¿Para qué tipo de pólipos sería más deseable un sistema CAD?		
Pólipos planos/deprimidos	87	89,7
Pólipos sésiles	3	3,1
Pólipos pedunculados	7	7,2
Según su experiencia, ¿cuál es el indicador más útil para establecer la malignidad de una lesión?		
Clasificación de París	21	23,3
Vascularidad	12	12,9
Superficie de la lesión (granularidad / no granularidad)	11	12,1
Patrón de criptas de Kudo	44	47,8
Los métodos existentes no son lo suficientemente reproducibles	31	34,1
¿Qué tipo de alerta preferiría para la detección de un pólipo?		
Alarma sonora	11	11,1
Señal visual en la pantalla (fuera de la imagen endoscópica)	43	43,4
Resaltar el pólipo en la imagen endoscópica (visor aumentado)	72	72,7
¿Cómo le gustaría recibir la información de sugerencia de diagnóstico?		
Alarma sonora	3	3,6
Indicador de semáforo (verde - benigno; amarillo - pre-maligno; rojo - maligno)	70	83,3
Texto	26	31,0
Resecaría completamente lesiones no polipoides con mayor confianza si dispusiera de un sistema CAD	79	86,8
El sistema CAD podría contribuir a detectar lesiones residuales de una cicatriz de una resección por partes	84	92,3

resaltando la zona correspondiente al pólipo dentro de la imagen endoscópica.

3. **Necesidades operacionales/físicas.** Los endoscopistas demandan un equipamiento con mejor calidad de imagen, ya que no siempre están disponibles sistemas HD.

5.2 Aprendizaje profundo para encontrar pólipos en colonoscopia: una revisión sistemática de la literatura

En este artículo se presenta una revisión sistemática del estado del arte sobre métodos de DL para detección, localización y segmentación de pólipos colorrectales. Para ello, se ha seguido la metodología PRISMA y se ha utilizado la cadena de búsqueda

((“colon” OR “colorectal”) AND (“cancer” OR “polyp”) AND (“deep learning”) AND (“detection” OR “localization” OR “segmentation”))

en las bases de datos ACM Digital Library, IEEE Digital Library, Web of Science, PubMed, Science@Direct, Scopus y Springer Link. Estas búsquedas arrojaron 1.332 resultados que se revisaron individualmente para comprobar si cumplían los criterios de inclusión. Finalmente, se analizó el texto completo de 35 artículos para la extracción de datos. En base a toda la información recopilada, las contribuciones de esta revisión sistemática son:

1. Análisis de conjuntos de datos públicos de imágenes colorrectales.
2. Análisis de métodos basados en DL para detección, localización y segmentación de pólipos.
3. Análisis de métricas empleadas.
4. Identificación de retos y recomendaciones para la aplicación de métodos de DL para pólipos colorrectales.

5.2.1 Análisis de conjuntos de datos públicos de imágenes colorrectales

En los trabajos analizados se identificaron siete conjuntos de datos de acceso público. En todos los casos, la verdad terreno es una máscara binaria indicando la zona que es pólipo dentro de la imagen. Asímismo, la mayoría de los conjuntos no proporcionan información clínica de las lesiones recogidas. También se caracterizan por tener un volumen limitado, lo que hace que sea común en los métodos analizados el uso de estrategias de aumento de datos para solventar esta carencia mediante el incremento artificial del conjunto de entrenamiento. Además, se ha identificado que no hay una metodología estándar para seleccionar las transformaciones y/o rangos a aplicar.

5.2.2 Análisis de métodos basados en aprendizaje profundo para detección, localización y segmentación de pólipos

Las redes se han clasificado en (1) CNN, tales como AlexNet, VGG16, VGG19 o GoogLeNet, incluyendo redes residuales como ResNet50; (2) FCN, como SegNet o U-Net; (3) GAN y redes neuronales recurrentes (RNN, de sus siglas en inglés *Recurret Neural Network*), incluyendo redes de gran memoria de corto plazo (LSTM, de sus siglas en inglés *Long Short-Term Memor*y). El uso de las CNN sobresale por encima del resto de tipo de redes en las tres tareas consideradas (Figura 5.1).

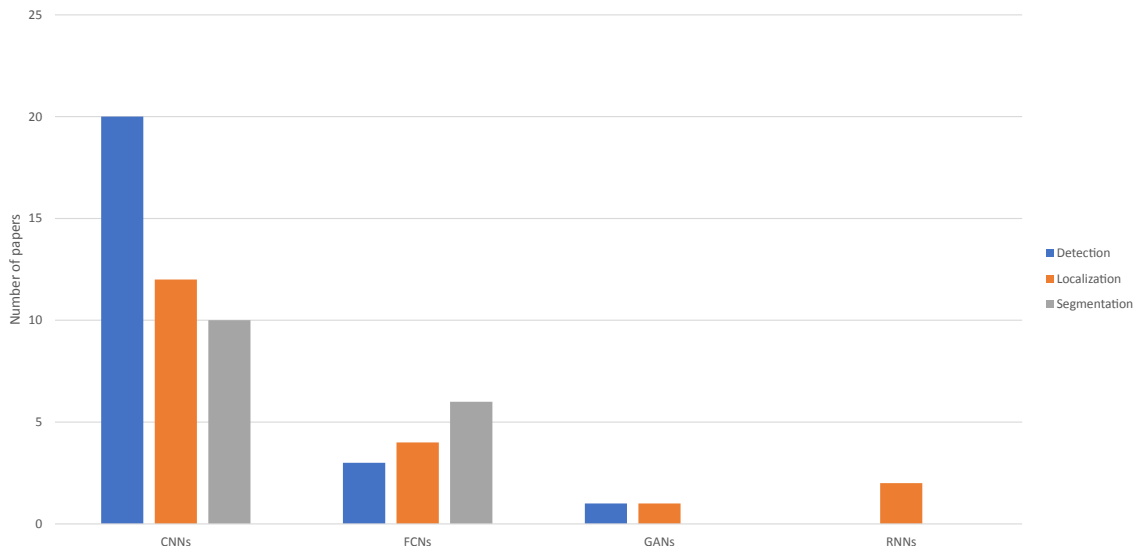


Figura 5.1: Redes para cada una de las tareas. CNN: redes neuronales convolucionales; FCN: redes completamente convolucionales; GAN: redes adversarias generativas; RNN: redes neuronales recurrentes [50].

Como clasificación principal se tomó la definida por Bernal *et al.* [111], en la que los trabajos se categorizaron como extremo a extremo, si únicamente se usa DL, e híbridos, si el modelo de DL se usa en combinación con métodos manuales. En los trabajos incluidos en la revisión sistemática, los métodos extremo a extremo e híbridos tienen una presencia similar en las tareas de detección y localización, pero son los primeros los que predominan para la tarea de segmentación.

Además, se estableció una clasificación secundaria dependiendo del uso que se haga de la red de DL (Figura 5.2):

1. Extractor de características, si la red de DL se usa para obtener el vector de características de manera automática en lugar de obtenerlas manualmente. Este vector se emplea posteriormente como entrada de un clasificador clásico.
2. Clasificación, si la red de DL se usa para etiquetar una imagen como pólipo o no pólipo.
3. Basado en recortes, si el método usa porciones de la imagen que se clasifican como pólipo o no pólipo. Además, sería posible extraer la localización de la lesión en función de la localización de los recortes.
4. Recuadro delimitador, si el método indica la localización del pólipo mediante un rectángulo, generalmente a partir de las coordenadas de la esquina superior izquierda, ancho y alto del mismo, usando una capa de regresión.
5. Segmentación semántica, si el método etiqueta cada píxel de la imagen como pólipo o no pólipo, generalmente mediante una arquitectura de tipo codificador-decodificador.

En este caso, para la tarea de detección, se encontraron mayoritariamente métodos de clasificación y basados en recortes. Por el contrario, los recuadros delimitadores y segmentación semántica son los tipos de modelos preferidos para localización y segmentación, respectivamente. Las combinaciones más habituales son métodos híbridos basados en recortes y métodos extremo a extremo de segmentación semántica (Figura 5.3).

La Tabla 5.2 recoje las principales ventajas y desventajas de las categorías establecidas para los métodos de DL.

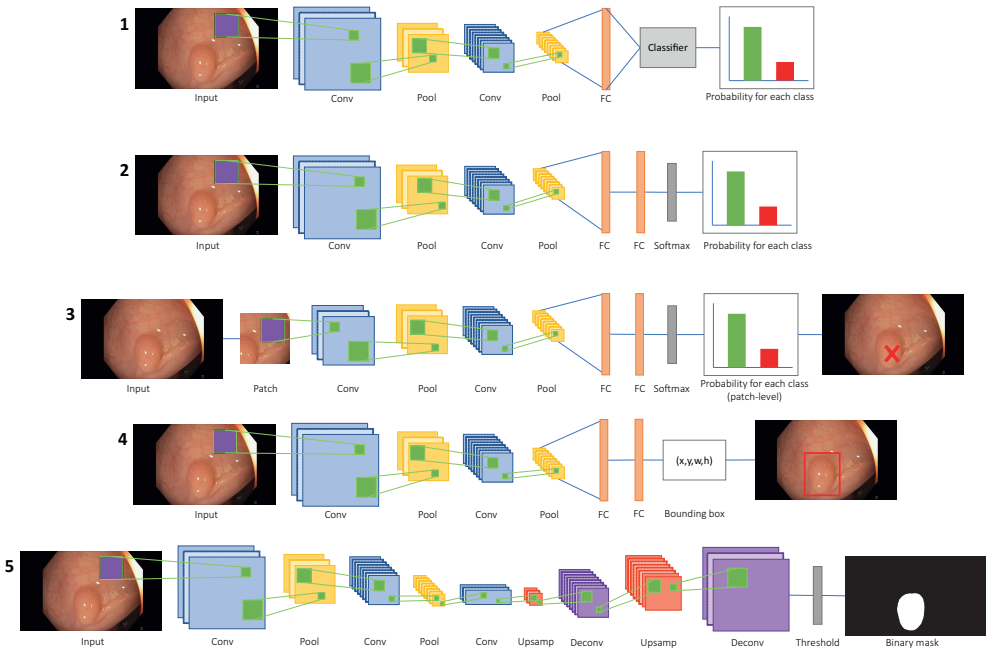


Figura 5.2: Representación esquemática de los cinco enfoques para métodos de DL dentro de la clasificación secundaria. De arriba a abajo: (1) extractor de características, (2) clasificación, (3) basado en recortes, (4) recuadro delimitador, y (5) segmentación semántica. Cada tipo de capa se representa con un color diferente: capa convolucional (conv), capa de muestreo (pool), capa completamente conectada (FC), capa de sobremuestreo (upsamp) y capa de convolución transpuesta (deconv). El campo receptivo se indica con un recuadro verde [50].

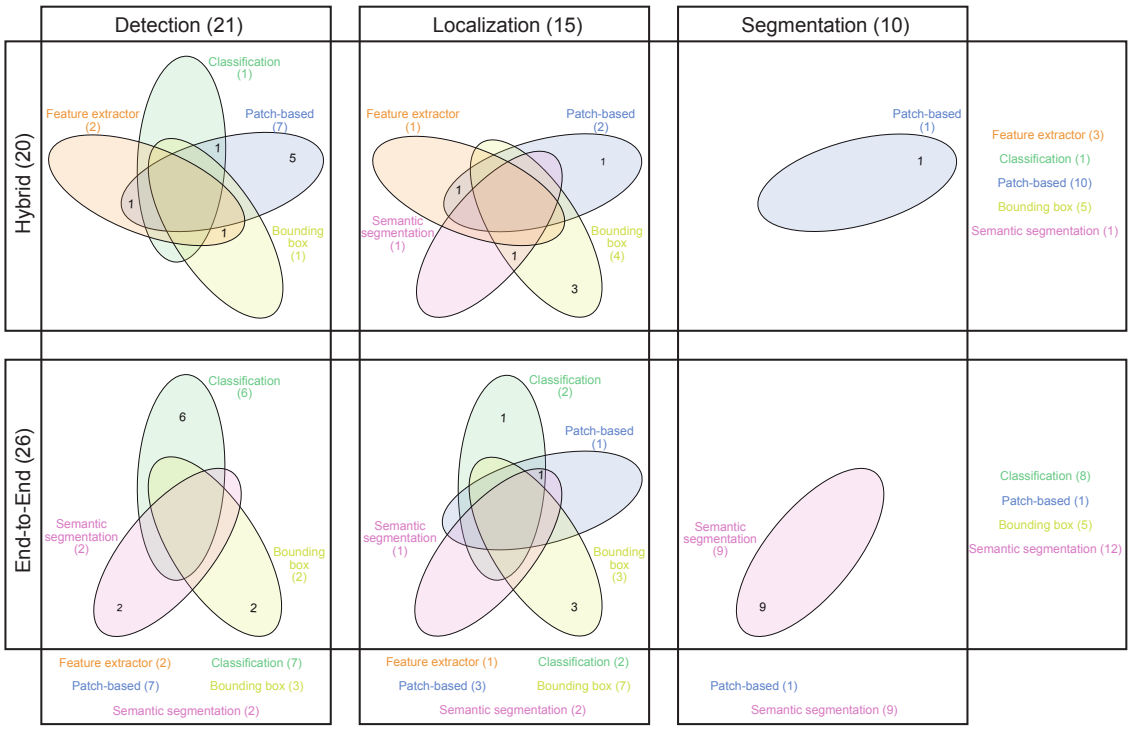


Figura 5.3: Categorización de trabajos por tarea [50].

Tabla 5.2: Ventajas y desventajas de las diferentes categorías de modelos de DL [50].

Categoría	Ventajas	Desventajas
Extremo a extremo	1. Aprendizaje automático de características relevantes 2. Mejor rendimiento que los métodos manuales 3. Alto grado de automatización	1. Requiere un conjunto de datos grande para el entrenamiento 2. La selección de la red, conjunto de datos e hiperparámetros puede tener gran influencia en el rendimiento
Híbrido	1. Combina información útil de características extraídas manualmente y bien conocidas 2. Más conveniente para conjuntos de datos pequeños	1. Selección de características según la experiencia y conocimiento del investigador
Extractor de características	1. Combina información útil de características extraídas manualmente y bien conocidas 2. El uso de redes preentrenadas sin ajuste de precisión no requiere datos etiquetados	1. No adapta la red al conjunto de datos objetivo, lo que potencialmente puede llevar a una transferencia de resultados subóptima o incluso negativa si el cambio de dominio es grande
Clasificación	1. Requiere el etiquetado más fácil (pólipo/no pólipo por imagen)	1. Requiere considerablemente más datos para converger
Basado en recortes	1. Incrementa el tamaño del conjunto de entrenamiento mediante la obtención de múltiples recortes de una única imagen	1. Método lento en general
Recuadro delimitador	1. Proporciona suficiente información al clínico, centrando su atención en un área sospechosa 2. Más velocidad computacional y coherencia de clasificación entre recortes 3. Capaz de converger sin etiquetas a nivel de píxel al mismo tiempo que predice información de localización 4. La anotación de recuadro delimitador es más fácil que a nivel de píxel 5. Tipo preferido para las tareas de localización	1. No proporciona un etiquetado a nivel de píxel 2. La densidad de objetos puede producir el solape de los recuadros
Segmentación semántica	1. Tipo preferido para las tareas de segmentación	1. Puede ser difícil obtener bordes precisos

5.2.3 Análisis de métricas empleadas

Tras analizar las métricas utilizadas en los trabajos de la revisión sistemática, no existe un estándar a la hora de presentar los resultados de los modelos de DL y para una misma tarea, cada trabajo calcula las métricas que considera oportunas. En la mayoría de los casos, las métricas provienen de los elementos de la matriz de confusión, calculadas bien a nivel de imagen o a nivel de píxel. De manera general, la exhaustividad es la más utilizada, seguida de precisión, exactitud y valor F1. Considerando las tareas, la exhaustividad a nivel de imagen es la métrica más utilizada para la detección y localización, mientras que el índice Jaccard es la más utilizada para segmentación, seguida de la exactitud a nivel de píxel.

5.2.4 Identificación de retos y recomendaciones para la aplicación de métodos de aprendizaje profundo para pólipos colorrectales

A la luz de los resultados obtenidos en la revisión sistemática, los retos y recomendaciones van en gran medida alineados con la falta de reproducibilidad de los trabajos. En primer lugar, es necesario establecer un conjunto de datos grande, con especial presencia de lesiones más complicadas de detectar, como pueden ser los pólipos planos 0-IIb según la clasificación de París. Además, debe establecerse claramente el conjunto de testeo, para que todos los métodos proporcionen los resultados sobre el mismo conjunto de imágenes, y las métricas a utilizar, en función de la verdad terreno y la tarea a realizar. En cuanto a las estrategias de aumento de datos, sigue siendo necesario identificar las mejores transformaciones y rangos, así como la inclusión de particularidades de la imagen endoscópica, como pueden ser las luces especulares o la falta de uniformidad. Por último, sería recomendable extender el uso de arquitecturas más allá de las CNN.

5.3 Efecto de transformaciones para aumento de datos en segmentación de pólipos

Este tercer artículo analiza los efectos de diferentes transformaciones para llevar a cabo un aumento de datos que compense el tamaño limitado de dos conjuntos de datos de acceso público: CVC-EndoSceneStill y Kvasir-SEG. Para ello, se ha entrenado una arquitectura basada en U-Net sin ninguna estrategia de aumento de datos, para establecer el nivel de referencia. Posteriormente, la misma red se ha entrenado aplicando diferentes transformaciones y rangos. Las transformaciones aplicadas (Figura 5.4) se agrupan en tres categorías:

1. **Transformaciones a nivel de imagen.** Engloban las transformaciones que se aplican a la imagen y a la máscara para modificarlas del mismo modo: desplazamiento en el ancho y alto, rotación, cizallamiento, aumento, reducción, volteo y deformación elástica.
2. **Transformaciones a nivel de píxel.** Estas transformaciones modifican los valores de los píxeles de la imagen, bien de todos los canales por igual o de cada canal de manera independiente. Incluye cambios en el brillo y contraste.
3. **Transformaciones basadas en la aplicación.** Este último grupo incluye características concretas de la aplicación, en este caso, la inclusión de luces especulares y la presencia de imágenes borrosas (simulado por un filtro de media).

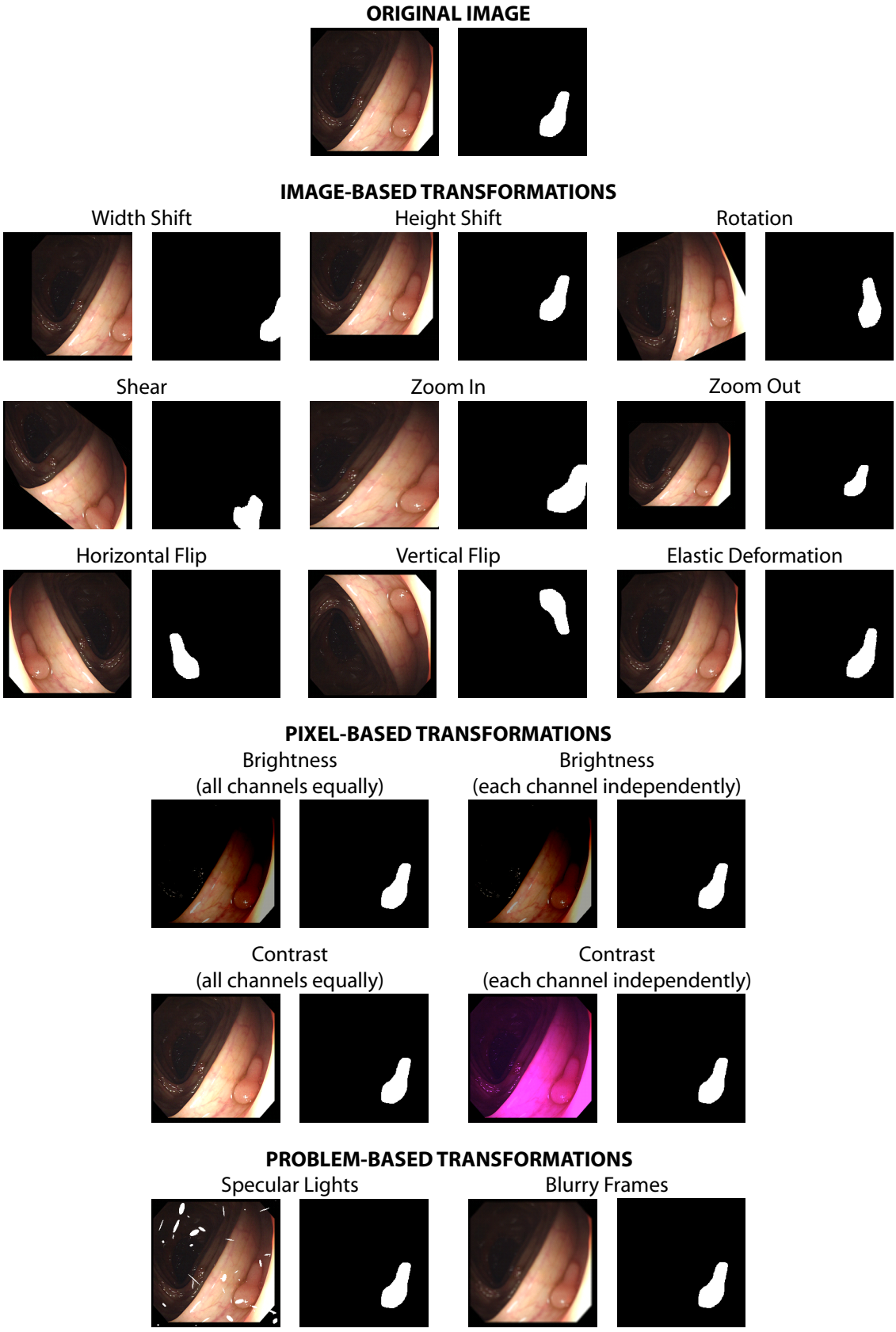


Figura 5.4: Imagen original e imágenes transformadas [138].

Para minimizar el efecto de aleatoriedad de las transformaciones dentro de los rangos establecidos, se llevaron a cabo diez repeticiones de entrenamiento por cada configuración de transformación y rango. Los resultados del nivel de referencia y de cada configuración se compararon mediante un test de permutación. Las transformaciones a nivel de imagen mostraron diferentes comportamientos dependiendo del conjunto de datos, de la transformación y del rango. En relación a las transformaciones a nivel de pixel, mejoran los resultados tanto modificando el brillo como el contraste, si bien estas mejoras son significativas sólo en el caso de CVC-EndoSceneStill. Finalmente, en cuanto a las transformaciones basadas en características concretas de la aplicación, la inclusión de luces especulares mejora levemente los resultados en los dos conjuntos de datos, mientras que las imágenes borrosas lo empeoran, de manera significativa únicamente en CVC-EndoSceneStill.

Se analizaron también los resultados de combinar diferentes transformaciones: la mejor transformación y rango para cada una de las tres categorías, el rango de las transformaciones a nivel de imagen con mejor resultado, siempre y cuando mejoren el valor de referencia, y las dos mejores transformaciones y rangos. Para CVC-EndoSceneStill, el valor medio se mantiene pero se minimiza la desviación. Por otro lado, la combinación de todas las transformadas a nivel de imagen no obtiene mejores resultados que combinar sólo dos transformaciones.

En vista de los resultados obtenidos, las transformaciones se pueden agrupar según su efecto en:

1. Transformadas que siempre mejoran el resultado en CVC-EndoSceneStill y Kvasir-SEG: volteo vertical, cambios en el brillo para cada canal independientemente, cambios en el contraste (cada canal independientemente y todos los canales por igual) e inclusión de luces especulares. Aunque se mejoran los resultados del nivel de referencia, la significancia se encuentra principalmente en las modificaciones del brillo y contraste en CVC-EndoSceneStill.
2. Transformaciones que siempre empeoran el resultado en CVC-EndoSceneStill y Kvasir-SEG: deformación elástica e imágenes borrosas. Mientras que la primera podría esperarse que mejorase el resultado, transformar las imágenes reduciendo su nitidez explica el empeoramiento de los resultados, no aportando beneficios en el aumento de datos.
3. Transformaciones cuyo resultado depende del rango seleccionado en CVC-EndoSceneStill y Kvasir-SEG: desplazamientos en alto y ancho, aumento y reducción. Los desplazamientos por encima del 40 % de la imagen mejoran los resultados, al igual que rangos pequeños en el aumento y rangos grandes en la reducción. Por tanto, es recomendable comprobar la adecuación del rango seleccionado cuando se apliquen estas transformaciones en la estrategia de aumento de datos.
4. Transformaciones cuyo resultado depende del conjunto de datos utilizado: rotación, cizallamiento, cambios en el brillo para todos los canales por igual y, en menor medida, volteo horizontal. Estas diferencias pueden deberse a las diferencias en el tamaño de pólipo, área válida, brillo y contraste de los dos conjuntos de datos considerados.

Por tanto, este trabajo pone de manifiesto que la selección de transformaciones y rangos tiene un impacto en los resultados del modelo y que, a pesar de no ser muy frecuentes, las transformaciones a nivel de píxel tienen gran potencial para mejorar la segmentación de pólipos colorrectales.

5.4 Eigenloss: función de pérdida combinada basada en PCA para segmentación de pólipos

Este trabajo presenta la función de pérdida *eigenloss* para segmentación de pólipos, obtenida como una combinación lineal de funciones de pérdida basada en un análisis de componentes principales (PCA, de sus siglas en inglés *Principal Component Analysis*). La disparidad de métricas empleadas, así como los diferentes tipos de funciones de pérdida, pone de manifiesto la gran cantidad de alternativas disponibles a la hora de implementar un modelo de DL. En este trabajo se analiza la influencia de varias funciones de pérdida individuales en varios modelos de DL para poder determinar si una descomposición basada en PCA permite obtener los coeficientes de una función de pérdida no redundante que dé lugar a mejores resultados que las funciones individuales u otros métodos de agregación básicos, como la suma o la media.

Para ello, se ha empleado CVC-EndoSceneStill como conjunto de datos. Además, se han seleccionado dos redes troncales y dos arquitecturas codificador-decodificador. Las redes troncales se diferencian en el tipo de bloque que se emplea para obtener el vector de características. Mientras que VGG-16 utiliza bloques convolucionales compuestos de capas CONV y de activación seguidas de una capa de MAXPOOL, Densenet121 hace uso de bloques densos, donde todas las capas están conectadas entre sí. Por otro lado, las arquitecturas codificador-decodificador difieren en cómo se integra la información del codificador en el decodificador: U-Net concatena los mapas de características del codificador en su nivel correspondiente del decodificador y LinkNet por el contrario suma la entrada de cada capa del codificador a la salida de la capa correspondiente en el decodificador. Por tanto, combinando las dos redes troncales y las dos arquitecturas codificador-decodificador obtenemos los cuatro modelos de DL considerados en este trabajo: U-Net-VGG-16 (Figura 5.5 A,C), U-Net-Densenet121 (Figura 5.5 B,C), LinkNet-VGG-16 (Figura 5.5 A,D) and LinkNet-Densenet121 (Figura 5.5 B,D).

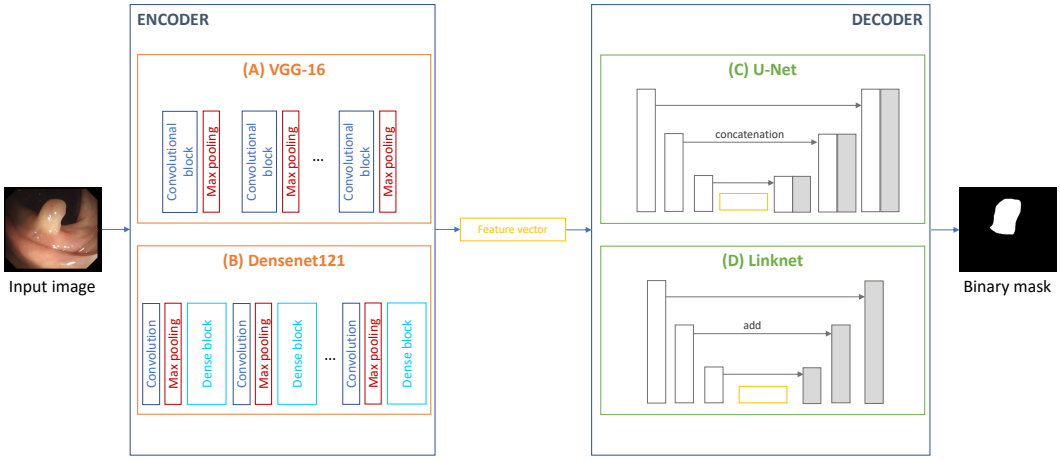


Figura 5.5: Redes troncales (A,B) y arquitecturas codificador-decodificador (C,D) [139].

Siendo $\mathbf{Y}$ la máscara de la verdad terreno e $\hat{\mathbf{Y}}$ la máscara de predicción, las funciones de pérdida individuales consideradas en este trabajo son:

1. Jaccard, $L_{Jaccard}(\mathbf{Y}, \hat{\mathbf{Y}})$. Indica el ratio de intersección de la verdad terreno y la máscara de predicción sobre la unión de ambas.

2. Dice, $L_{Dice}(\mathbf{Y}, \hat{\mathbf{Y}})$. Mide las similitudes entre dos regiones.
3. Entropía binaria cruzada, $L_{BEC}(\mathbf{Y}, \hat{\mathbf{Y}})$. Maximiza la exactitud de la máscara de predicción considerando tanto la clase pólipo como la clase fondo.
4. Focal binaria, $L_{BF}(\mathbf{Y}, \hat{\mathbf{Y}})$. Similar a la anterior, pero contrarrestando el efecto de clases desequilibradas.
5. Tversky, $L_{Tversky}(\mathbf{Y}, \hat{\mathbf{Y}})$. Controla la penalización de FP y FN.
6. Focal Tversky, $L_{FocalTversky}(\mathbf{Y}, \hat{\mathbf{Y}})$. Similar a la anterior, pero contrarrestando el efecto de clases desequilibradas.
7. Lovász-Hinge, $L_{Lovasz-Hinge}(\mathbf{Y}, \hat{\mathbf{Y}})$. Basada en la extensión convexa de Lovász de pérdida sub-modular.

Cada modelo se ha entrenado con cada una de estas siete funciones de pérdida y para cada imagen del conjunto de testeo, se calculan las siete métricas definidas en la Sección 4.4: exactitud, precisión, exhaustividad, especificidad, valor F2, índice Jaccard e índice Dice. La Tabla 5.3 recoge los resultados de la métrica Jaccard de cada uno de los modelos. El resto de métricas pueden encontrarse en el texto completo del artículo (Sección 8.4).

Tabla 5.3: Resultados de Jaccard para los diferentes modelos y funciones de pérdida. Se indica en negrita el mejor resultado de cada modelo [139].

Función de pérdida	U-Net-VGG-16	U-Net-Densenet121	LinkNet-VGG-16	LinkNet-Densenet121
Jaccard	$59,77 \pm 35,03$	$72,17 \pm 31,06$	$62,75 \pm 34,19$	$71,57 \pm 28,02$
Dice	$56,38 \pm 34,59$	$69,88 \pm 32,63$	$62,53 \pm 35,45$	$69,71 \pm 32,18$
Entropía Binaria Cruzada	$59,48 \pm 36,35$	$68,07 \pm 31,89$	$62,65 \pm 33,05$	$68,74 \pm 32,73$
Focal Binaria	$60,49 \pm 33,66$	$70,01 \pm 30,18$	$58,36 \pm 34,85$	$69,11 \pm 32,62$
Tversky	$55,34 \pm 36,12$	$73,19 \pm 30,17$	$63,14 \pm 34,49$	$70,92 \pm 30,11$
Focal Tversky	$57,20 \pm 35,57$	$\mathbf{75,10 \pm 31,24}$	$62,81 \pm 34,66$	$70,84 \pm 28,83$
Lovász-Hinge	$38,23 \pm 37,64$	$72,26 \pm 30,62$	$59,60 \pm 34,49$	$\mathbf{72,41 \pm 29,88}$
Suma	$53,58 \pm 37,44$	$72,89 \pm 29,77$	$60,40 \pm 34,20$	$68,36 \pm 32,11$
Media	$\mathbf{62,83 \pm 33,45}$	$70,75 \pm 30,84$	$62,77 \pm 34,16$	$67,77 \pm 32,82$
Eigenloss	$59,66 \pm 35,32$	$69,35 \pm 31,55$	$\mathbf{63,24 \pm 33,86}$	$71,84 \pm 29,25$
Eigenloss normalizada	$56,56 \pm 33,99$	$71,86 \pm 31,02$	$60,62 \pm 35,21$	$71,57 \pm 31,28$

Tras comprobar la viabilidad de aplicar la PCA a los datos obtenidos en cada modelo, este análisis identificó una componente independiente para cada uno (Tabla 5.4) y por tanto, se define la *eigenloss* para esta componente principal (Ecuación 5.1) como una combinación lineal de las funciones de pérdida individuales, tomando como coeficientes los eigenvalores obtenidos (valores α_{1n} para cada modelo en la Tabla 5.4).

$$\begin{aligned} L_{1Eigenloss}(\mathbf{Y}, \hat{\mathbf{Y}}) = & \alpha_{1Jaccard} \cdot L_{Jaccard}(\mathbf{Y}, \hat{\mathbf{Y}}) + \alpha_{1Dice} \cdot L_{Dice}(\mathbf{Y}, \hat{\mathbf{Y}}) + \alpha_{1BCE} \cdot L_{BCE}(\mathbf{Y}, \hat{\mathbf{Y}}) + \\ & \alpha_{1BF} \cdot L_{BF}(\mathbf{Y}, \hat{\mathbf{Y}}) + \alpha_{1Tversky} \cdot L_{Tversky}(\mathbf{Y}, \hat{\mathbf{Y}}) + \\ & \alpha_{1FocalTversky} \cdot L_{FocalTversky}(\mathbf{Y}, \hat{\mathbf{Y}}) + \\ & \alpha_{1Lovasz-Hinge} \cdot L_{Lovasz-Hinge}(\mathbf{Y}, \hat{\mathbf{Y}}) \end{aligned} \tag{5.1}$$

Los resultados de los modelos entrenados con la *eigenloss* se incluyen en la Tabla 5.3, junto con los de otras combinaciones lineales, tales como la suma y la media, así como la *eigenloss*

normalizada, en la que los coeficentes α_{1n} se normalizan según la fórmula

$$\alpha'_{1n} = \frac{\alpha_{1n}}{\sum_n \alpha_{1n}}$$

con $n \in Jaccard, Dice, BCE, BF, Tversky, FocalTversky, Lovasz - Hinge$.

Tabla 5.4: Resultados de la PCA. Com.: Comunidalidad, α_{1n}: eigenvalores [139].

	U-Net-VGG-16		U-Net-Densenet121		LinkNet-VGG-16		LinkNet-Densenet121	
Eigenvalor	5,803		6,076		5,913		5,998	
% Varianza	82,91		86,80		84,47		85,69	
Función de pérdida	Com.	α_{1n}	Com.	α_{1n}	Com.	α_{1n}	Com.	α_{1n}
Jaccard	0,875	0,935	0,819	0,905	0,843	0,918	0,819	0,905
Dice	0,845	0,919	0,894	0,946	0,864	0,930	0,878	0,937
Entropía Binaria Cruzada	0,835	0,914	0,785	0,886	0,866	0,930	0,842	0,918
Focal Binaria	0,826	0,909	0,892	0,944	0,805	0,897	0,861	0,928
Tversky	0,881	0,939	0,906	0,952	0,842	0,917	0,881	0,939
Focal Tversky	0,884	0,940	0,869	0,932	0,888	0,942	0,878	0,937
Lovász-Hinge	0,657	0,811	0,911	0,954	0,805	0,897	0,840	0,916

La *eigenloss* propuesta mejora los resultados a nivel general con respecto a las funciones de pérdida individuales en todos los modelos, ya que proporciona valores más altos en todas las métricas, a pesar de que el valor máximo de alguna de ellas pueda encontrarse en las funciones de pérdida individuales. Además, los resultados del valor F2 son más altos en la función de pérdida combinada, lo que implica que el balance entre precisión y exhaustividad es mejor en esta que en las funciones de pérdida individuales. Por último, la *eigenloss* minimiza la variabilidad en el índice Jaccard. A pesar de las mejoras mencionadas con respecto a las funciones de pérdida individuales, la *eigenloss* sólo es mejor que la suma o la media de las siete funciones cuando se usa la arquitectura LinkNet.

Los resultados obtenidos se pueden comparar de manera directa con los trabajos de otros autores que han utilizado también el conjunto de datos de CVC-EndoSceneStill. La Tabla 5.5 muestra que la *eigenloss* propuesta obtiene mejores resultados en cuatro de las cinco métricas disponibles para comparación con otros trabajos. Además, cabe mencionar que la exactitud no es la métrica más adecuada en este tipo de problemas en los que existen un gran desbalanceo entre las clases positiva y negativa.

5.5 Conjunto de datos de PICCOLO: comparativa de modelos y conjuntos de datos

Este quinto artículo presenta el conjunto de datos PICCOLO, que contiene imágenes de luz blanca y NBI, junto con una comparativa de diferentes modelos de DL y conjuntos de datos. Actualmente este es el mayor conjunto de datos de acceso público con imágenes manualmente segmentadas por gastroenterólogos expertos. Además, proporciona la distribución en los diferentes subconjuntos de entrenamiento, validación y testeo para asegurar la independecia entre ellos así como facilitar la comparación de modelos de DL que lo utilicen. Por último, también

Tabla 5.5: Comparación de resultados de los cuatro modelos entrenados con la *eigenloss* con otros trabajos que también usan CVC-EndoSceneStill. Se indica en negrita el mejor resultado para cada métrica [139].

Trabajo	ACC	PREC	REC	SPEC
U-Net-VGG-16	$93,00 \pm 10,91$	$80,31 \pm 31,35$	$65,07 \pm 37,70$	$99,00 \pm 2,25$
U-Net-Densenet121	$94,43 \pm 9,78$	$\mathbf{89,83 \pm 21,63}$	$77,39 \pm 33,28$	$99,06 \pm 2,18$
LinkNet-VGG-16	$93,50 \pm 10,60$	$83,21 \pm 28,49$	$69,52 \pm 36,50$	$\mathbf{99,28 \pm 1,76}$
LinkNet-Densenet121	$94,59 \pm 9,61$	$88,65 \pm 18,25$	$\mathbf{80,23 \pm 30,78}$	$99,05 \pm 1,92$
Vázquez *et al.* [132]	$\mathbf{96,77}$	–	–	–
Wichakam *et al.* [107]	–	$88,84$	$78,14$	–
Wickstrøm *et al.* [140]	$94,90$	–	–	–

Trabajo	Valor F2	Índice Jaccard	Índice Dice
U-Net-VGG-16	$65,45 \pm 36,70$	$59,66 \pm 35,32$	$67,04 \pm 35,39$
U-Net-Densenet121	$76,41 \pm 31,97$	$69,35 \pm 31,55$	$76,33 \pm 30,39$
LinkNet-VGG-16	$69,51 \pm 35,31$	$63,24 \pm 33,86$	$70,61 \pm 33,68$
LinkNet-Densenet121	$\mathbf{79,18 \pm 29,26}$	$\mathbf{71,84 \pm 29,25}$	$\mathbf{79,07 \pm 27,07}$
Vázquez *et al.* [132]	–	$56,07$	–
Wichakam *et al.* [107]	–	$59,36$	$78,61$
Wickstrøm *et al.* [140]	–	$58,70$	–

proporciona metadatos clínicos de cada una de las lesiones, a diferencia de otros conjuntos de datos.

En primer lugar, se establece un protocolo de adquisición y anotación para poder obtener las imágenes manualmente anotadas por gastroenterólogos expertos. Estos protocolos se han diseñado de manera que no se interfiera con la práctica clínica. Finalmente, el conjunto de datos de PICCOLO contiene 76 lesiones de 40 pacientes diferentes. Para cada una de estas lesiones, se proporcionan una serie de metadatos clínicos (Apéndice A.4.1).

De las 76 lesiones, 62 lesiones incluyen imágenes de luz blanca y NBI, mientras que las restantes 14 sólo presentan imágenes de luz blanca. En total, se obtuvieron más de 145.000 fotogramas, de los que se seleccionaron 3.433 para ser anotados manualmente. Para cada fotograma, el conjunto de datos proporciona tres imágenes, el propio fotograma que muestra la imagen de luz blanca o NBI en un archivo .png, la máscara binaria que indica la zona correspondiente a la lesión, manualmente delimitada, y la máscara binaria *void* que indica el área negra de la imagen original (Figura 5.6).

Las imágenes anotadas se han distribuido en conjuntos de entrenamiento, validación y testeo. Con el fin de asegurar la independencia entre ellos, las imágenes procedentes del mismo paciente se han asignado al mismo conjunto. La Tabla 5.6 muestra la distribución de imágenes según los metadatos clínicos y los conjuntos definidos.

El conjunto de datos de PICCOLO es de acceso público y puede solicitarse el acceso a través de la web del Biobanco Vasco. El uso del conjunto de datos está limitado a fines investigadores y educativos, mientras que el uso comercial está prohibido sin consentimiento escrito previo.

Con el fin de establecer la utilidad del conjunto de datos de PICCOLO, los modelos presentados en la Sección 5.4 se han entrenado haciendo uso de los tres conjuntos de datos disponibles: CVC-EndoSceneStill, Kvasir-SEG y PICCOLO. Para poder establecer comparaciones de manera adecuada, los resultados se proporcionan utilizando siempre las mismas imágenes de testeo (182, 200 y 333 para CVC-EndoSceneStill, Kvasir-SEG y PICCOLO, respectivamente) además de crear

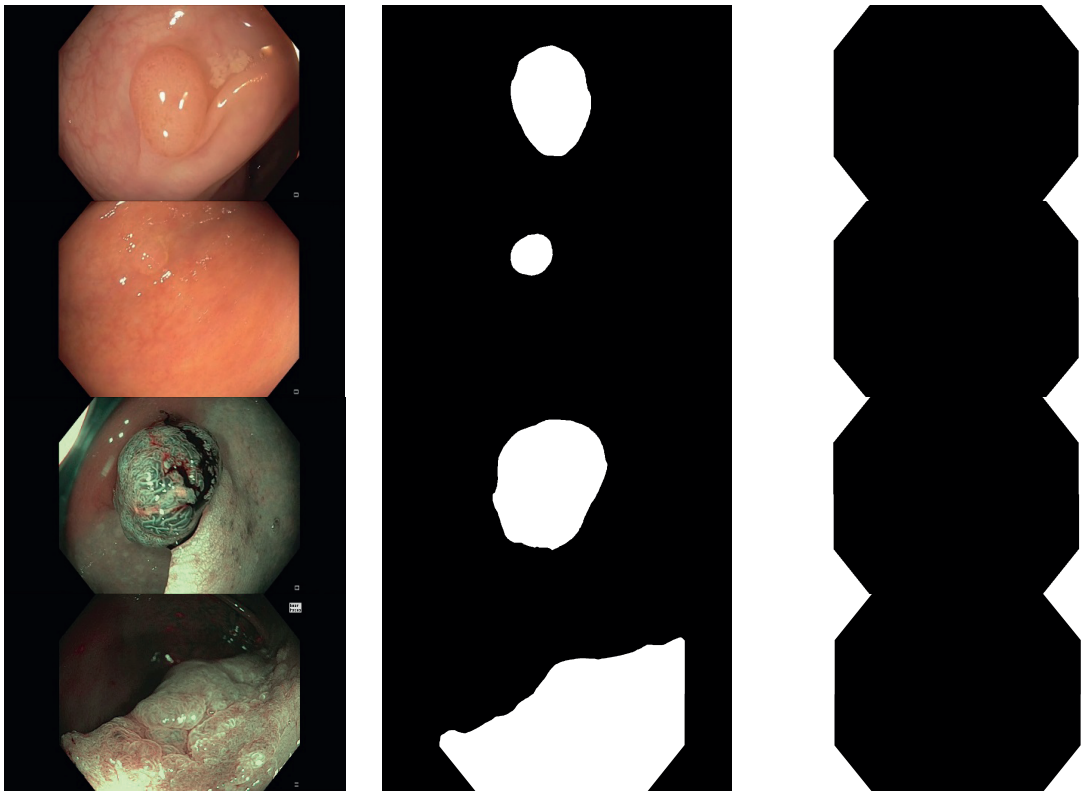


Figura 5.6: Ejemplos de las imágenes proporcionadas por PICCOLO [135].

un conjunto de testeo que las engloba todas. La Tabla 5.7 recoge los resultados del índice Jaccard, mientras que los resultados para el resto de métricas calculadas se proporcionan en el material suplementario correspondiente (Apéndices A.4.2, A.4.3, A.4.4 y A.4.5). Analizando los resultados de los diferentes conjuntos según se usen para entrenamiento y/o testeo, se puede concluir que PICCOLO da lugar a modelos más generalizados, que proporcionan resultados más uniformes para todos los conjuntos de datos, en lugar de obtener los mejores resultados sólo en sus propias imágenes de testeo. Del mismo modo, cabe mencionar que los resultados con PICCOLO son inferiores a los de los otros conjuntos de datos, lo que puede deberse a que contiene mayor variabilidad de pólipos, y por tanto siendo más difícil el aprendizaje. Por último, sólo el conjunto de datos de PICCOLO combina imágenes de luz blanca y NBI, por lo que podría aumentar la robustez frente a cambios de color, dando lugar a esa mejor capacidad de generalización de los modelos.

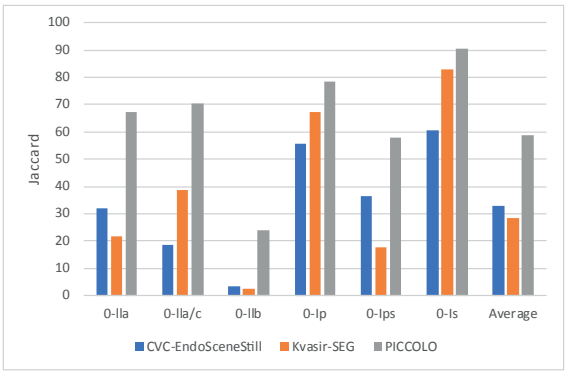
Tabla 5.6: Distribución de imágenes en el conjunto de datos de PICCOLO. N/A: no aplicable [135].

Categoría	Items	Núm. lesiones	Núm. frames WL	Núm. frames NBI	Conj. entrenamiento	Conj. validación	Conj. testeo
Tipo de imagen	WL	76	2132	-	1382	558	192
	NBI	62	-	1302	821	340	141
Clasificación de Paris	0-Ip	10	191	193	274	81	29
	0-Ips	7	193	107	245	41	14
	0-Is	16	387	253	433	176	31
	0-IIa	29	928	436	1052	263	49
	0-IIa/c	4	83	91	27	122	25
	0-IIb	2	85	21	-	48	58
	N/A	8	264	201	172	166	127
Clasificación NICE	Tipo 1	17	404	284	435	139	114
	Tipo 2	50	1454	782	1552	592	92
	Tipo 3	8	264	201	172	166	127
	N/A	1	9	35	44	-	-
Diagnóstico	Adenocarcinoma	8	264	201	172	166	127
	Adenoma	50	1454	782	1552	592	92
	Hiperplasia	17	404	284	435	139	114
	N/A	1	9	35	44	-	-
Estratificación histológica	Displasia de alto grado	12	409	168	360	217	-
	Hiperplasia	13	312	225	342	139	56
	Adenocarcinoma invasivo	8	264	201	172	166	127
	Displasia de bajo grado	1	13	16	-	-	29
	No displasia	41	1124	657	1285	375	121
	N/A	1	9	35	44	-	-

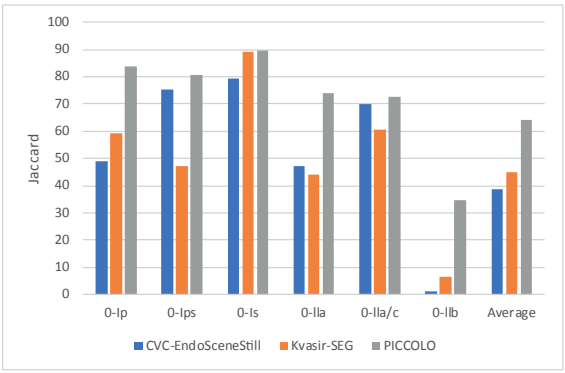
Tabla 5.7: Índice Jaccard para cada experimento. Se indica en negrita el mejor resultado para cada conjunto de datos de entrenamiento y validación.

Conjunto de datos para entrenamiento y validación	Modelo	Conjunto de datos para testeo			
		CVC-EndoSceneStill	Kvasir-SEG	PICCOLO	Todos
CVC-EndoSceneStill	U-Net-VGG-16	$56,80 \pm 36,09$	$46,29 \pm 29,36$	$32,82 \pm 34,81$	$42,69 \pm 35,13$
	U-Net-Densenet121	$67,31 \pm 34,47$	$56,12 \pm 34,29$	$38,76 \pm 39,05$	$50,88 \pm 38,51$
	LinkNet-VGG-16	$62,77 \pm 33,81$	$51,13 \pm 27,67$	$30,55 \pm 33,57$	$44,51 \pm 34,86$
	LinkNet-Densenet121	$\mathbf{72,16 \pm 30,93}$	$\mathbf{56,69 \pm 33,68}$	$\mathbf{39,52 \pm 37,90}$	$\mathbf{52,63 \pm 37,53}$
Kvasir-SEG	U-Net-VGG-16	$32,44 \pm 38,79$	$64,23 \pm 29,78$	$28,51 \pm 35,12$	$39,50 \pm 37,98$
	U-Net-Densenet121	$\mathbf{71,82 \pm 29,87}$	$74,13 \pm 23,40$	$44,78 \pm 38,73$	$\mathbf{59,87 \pm 35,72}$
	LinkNet-VGG-16	$58,62 \pm 36,18$	$72,53 \pm 23,92$	$40,43 \pm 36,41$	$54,04 \pm 35,99$
	LinkNet-Densenet121	$69,10 \pm 32,53$	$\mathbf{74,52 \pm 22,81}$	$\mathbf{44,92 \pm 37,37}$	$59,35 \pm 35,33$
PICCOLO	U-Net-VGG-16	$47,76 \pm 36,46$	$52,64 \pm 30,41$	$58,74 \pm 36,06$	$54,24 \pm 34,93$
	U-Net-Densenet121	$62,41 \pm 34,78$	$\mathbf{65,33 \pm 30,66}$	$\mathbf{64,01 \pm 36,23}$	$\mathbf{63,97 \pm 34,35}$
	LinkNet-VGG-16	$54,58 \pm 35,63$	$58,99 \pm 30,40$	$54,46 \pm 38,88$	$55,76 \pm 35,87$
	LinkNet-Densenet121	$\mathbf{64,18 \pm 33,04}$	$59,61 \pm 33,80$	$60,14 \pm 38,31$	$61,02 \pm 35,79$

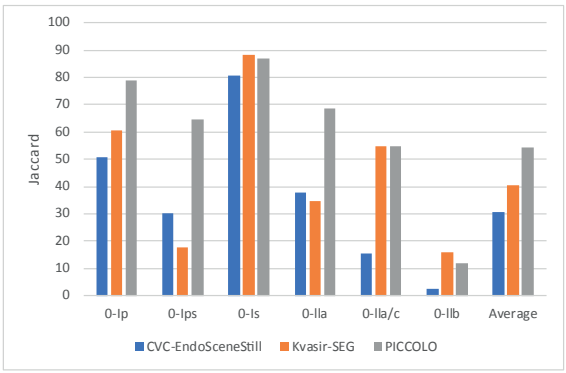
Si se analizan los resultados según la morfología de los pólipos (Figura 5.7), lo cual sólo es posible con PICCOLO ya que ni CVC-EndoSceneStill ni Kvasir-SEG proporcionan dicha información, los pólipos planos (0-IIb) obtienen los resultados más bajos en todos los modelos, mientras que los pólipos sésiles y pedunculados (0-Ip, 0-Ips y 0-Is) consiguen mejores resultados que los demás (0-IIa, 0-IIa/c y 0-IIb).



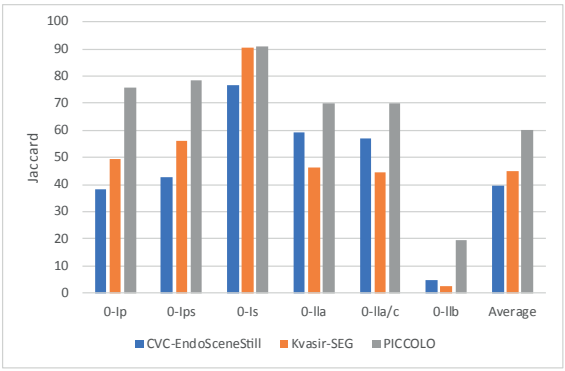
(a) U-Net-VGG-16



(b) U-Net-Densenet121



(c) LinkNet-VGG-16



(d) LinkNet-Densenet121

Figura 5.7: Índice Jaccard en el subconjunto de testeo del conjunto de datos de PICCOLO para cada uno de los modelos [135].

6 Conclusiones

En esta tesis se han analizado diferentes estrategias que proporcionan mejoras en la detección, localización y segmentación automática de polipos colorrectales basadas en técnicas de aprendizaje profundo. A la luz de los resultados expuestos en el Capítulo 5, las conclusiones de este trabajo de tesis se enumeran a continuación:

1. Los sistemas CAD deben poner mayor énfasis en aquellos pólipos colorrectales que son de pequeñas dimensiones y apariencia sutil.
2. En cuanto a los elementos necesarios para implementar modelos de aprendizaje profundo en esta aplicación, se concluye que:
   a) Los conjuntos de datos de pólipos colorrectales públicos disponibles antes de la realización de esta tesis tienen un número de imágenes escaso para técnicas de aprendizaje profundo, además de no proporcionar metadatos clínicos.
   b) No existe un consenso a la hora de elegir las transformaciones ni sus rangos para hacer aumento de datos.
   c) Las redes neuronales convolucionales son la arquitectura más utilizada, al mismo tiempo que los métodos extremo a extremo tienen mayor presencia que los híbridos.
   d) Aunque no existe consenso, la exhaustividad es la métrica más empleada para detección y localización, mientras que el índice Jaccard destaca para segmentación.
3. Con unos protocolos de adquisición y anotación que no interfieren con la práctica clínica se ha generado el conjunto de datos público de mayor volumen con anotaciones manuales que contiene 3.433 imágenes de pólipos colorrectales y sus metadatos clínicos.
4. Aunque las transformaciones a nivel de imagen son más utilizadas en el aumento de datos, las transformaciones a nivel de píxel presentan mayor potencial para mejorar los resultados de los modelos de aprendizaje profundo en este tipo de imágenes.
5. El uso de los elementos del eigenvector para generar la función de pérdida *eigenloss* junto con el modelo LinkNet-DenseNet121 proporciona métricas de evaluación más estables en este campo de aplicación.
6. Los modelos entrenados con el conjunto de datos generado presentan mayor capacidad de generalización que si se entrenan con otros conjuntos de datos de acceso público.

7 Bibliografía

[1] International Agency for Research on Cancer, *World cancer report 2014*, B. W. Stewart y C. P. Wild, Eds. International Agency for Research on Cancer, 2014. ISBN 978-92-832-0429-9

[2] International Agency for Research on Cancer, *World cancer report 2020*, C. P. Wild, E. Weiderpass, y B. W. Stewart, Eds. International Agency for Research on Cancer, 2020. ISBN 9789283204299

[3] J. Weitz, M. Koch, J. Debus, T. Höhler, P. R. Galle, y M. W. Büchler, "Colorectal cancer," *Lancet*, vol. 365, no. 9454, pp. 153–165, 2005. doi:10.1016/S0140-6736(05)17706-X

[4] R. Singh, L. Zorrón Cheng Tao Pu, D. Koay, y A. Burt, "Sessile serrated adenoma/polyps: Where are we at in 2016?" *World Journal of Gastroenterology*, vol. 22, no. 34, pp. 7754–7759, 2016. doi:10.3748/wjg.v22.i34.7754

[5] J. E. G. IJspeert *et al.*, "Detection rate of serrated polyps and serrated polyposis syndrome in colorectal cancer screening cohorts: a European overview," *Gut*, vol. 66, no. 7, pp. 1225–1232, 2017. doi:10.1136/gutjnl-2015-310784

[6] H. Sung *et al.*, "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, vol. 71, no. 3, pp. 209–249, 2021. doi:10.3322/caac.21660

[7] International Agency for Research on Cancer, "Cancer tomorrow," 2021. [Online]. Disponible en: https://gco.iarc.fr/tomorrow/en/dataviz/bars?mode=cancer&types=1&cancers=8&bar_mode=grouped&sexes=0&group_populations=1&multiple_cancers=1&sort_by=value0 Último acceso: 23 de julio de 2021.

[8] International Agency for Research on Cancer, "Cancer today," 2021. [Online]. Disponible en: https://gco.iarc.fr/today/online-analysis-map?v=2020&mode=population&mode_population=continents&population=900&populations=900&key=asr&sex=0&cancer=41&type=0&statistic=5&prevalence=0&population_group=0&ages_group%5B%5D=0&ages_group%5B%5D=17&nb_items=10&group_cancer=1&include_nmsc=1&include_nmsc_other=1&projection=natural-earth&color_palette=default&map_scale=quantile&map_nb_colors=5&continent=5&show_ranking=0&rotate=%255B10%252C0%255D Último acceso: 23 de julio de 2021.

[9] International Agency for Research on Cancer, "Cancer today," 2021. [Online]. Disponible en: https://gco.iarc.fr/today/online-analysis-map?v=2020&mode=population&mode_population=continents&population=900&populations=900&key=asr&sex=0&cancer=41&

type=1&statistic=5&prevalence=0&population_group=0&ages_group%5B%5D=0&ages_group%5B%5D=17&nb_items=10&group_cancer=1&include_nmsc=1&include_nmsc_other=1&projection=natural-earth&color_palette=default&map_scale=quantile&map_nb_colors=5&continent=5&show_ranking=0&rotate=%255B10%252C0%255D Último acceso: 23 de julio de 2021.

[10] S. Rasmussen *et al.*, "Predictive values of colorectal cancer alarm symptoms in the general population: a nationwide cohort study," *British Journal of Cancer*, vol. 120, no. 6, pp. 595–600, 2019. doi:10.1038/s41416-019-0385-x

[11] A. Wiegering *et al.*, "Improved survival of patients with colon cancer detected by screening colonoscopy," *International Journal of Colorectal Disease*, vol. 31, no. 5, pp. 1039–1045, 2016. doi:10.1007/s00384-015-2501-6

[12] C. C. Moreno *et al.*, "Colorectal cancer initial diagnosis: screening colonoscopy, diagnostic colonoscopy, or emergent surgery, and tumor stage and size at initial presentation," *Clinical Colorectal Cancer*, vol. 15, no. 1, pp. 67–73, 2016. doi:10.1016/j.clcc.2015.07.004

[13] I. Ruiz Martín, C. Delgado Fernández, y S. Fernández Fernández, "La detección precoz sintomática. Signos de alarma generales de cáncer," en *Manual SEOM de prevención y diagnóstico precoz del cáncer*, Sociedad Española de Oncología Médica, Ed., 2017. ISBN 978-84-697-4630-1

[14] I. Chirivella González, M. T. Martínez Martínes, y A. Viala Monleón, "Detección precoz del cáncer. Programas de cribado," en *Manual SEOM de prevención y diagnóstico precoz del cáncer*, Sociedad Española de Oncología Médica, Ed., 2017. ISBN 978-84-697-4630-1

[15] P. Basu *et al.*, "Status of implementation and organization of cancer screening in The European Union Member States—Summary results from the second European screening report," *International Journal of Cancer*, vol. 142, no. 1, pp. 44–56, 2018. doi:10.1002/ijc.31043

[16] E. Ferlizza, R. Solmi, M. Sgarzi, L. Ricciardiello, y M. Lauriola, "The roadmap of colorectal cancer screening," *Cancers*, vol. 13, no. 5, p. 1101, 2021. doi:10.3390/cancers13051101

[17] J. Brunet Vidal, G. Binefa Rodríguez, y M. Obón Santacana, "Prevención del cáncer colorrectal," en *Manual SEOM de prevención y diagnóstico precoz del cáncer*, Sociedad Española de Oncología Médica, Ed., 2017. ISBN 978-84-697-4630-1

[18] Junta de Extremadura, *Plan integral contra el cáncer en Extremadura 2017-2021*, 2017. [Online]. Disponible en: https://saludextremadura.ses.es/filescms/web/uploaded_files/CustomContentResources/PLAN%20INTEGRAL%20CONTRA%20EL%20CANCER%202017-2021.pdf Último acceso: 23 de julio de 2021.

[19] T. R. Levin *et al.*, "Effects of organized colorectal cancer screening on cancer incidence and mortality in a large, community-based population," *Gastroenterology*, vol. 18, pp. 34 783–34 788, 2018. doi:10.1053/j.gastro.2018.07.017

[20] J. Mar *et al.*, "Valoración del coste económico del cáncer colorrectal según estadio tumoral," *Cirugía Española*, vol. 95, no. 2, pp. 89–96, 2017. doi:10.1016/j.ciresp.2017.01.001

[21] C. B. Williams, "Insertion technique," en *Colonoscopy. Principles and Practice*, J. D. Waye, D. K. Rex, y C. B. Williams, Eds. Blackwell Publishing Ltd, 2005. ISBN 10 1-4051-1449-5

[22] E. Arana-Arri *et al.*, "Screening colonoscopy and risk of adverse events among individuals undergoing fecal immunochemical testing in a population-based program: A nested case-control study," *United European Gastroenterology Journal*, vol. 6, no. 5, pp. 755–764, 2018. doi:10.1177/2050640618756105

[23] J. P. Berros Fombella, S. Aguilar Huergo, y P. García Teijido, "Enfermedades premalignas," en *Manual SEOM de prevención y diagnóstico precoz del cáncer*, Sociedad Española de Oncología Médica, Ed., 2017. ISBN 978-84-697-4630-1

[24] D. K. Rex, "Polyp detection at colonoscopy: Endoscopist and technical factors," *Best Practice and Research: Clinical Gastroenterology*, vol. 31, no. 4, pp. 425–433, 2017. doi:10.1016/j.bpg.2017.05.010

[25] M. Ferlitsch *et al.*, "Colorectal polypectomy and endoscopic mucosal resection (EMR): European Society of Gastrointestinal Endoscopy (ESGE) clinical guideline," *Endoscopy*, vol. 49, no. 3, pp. 270–297, 2017. doi:10.1055/s-0043-102569

[26] M. Lund, M. Trads, S. H. Njor, R. Erichsen, y B. Andersen, "Quality indicators for screening colonoscopy and colonoscopist performance and the subsequent risk of interval colorectal cancer: a systematic review," *JBI Database of Systematic Reviews and Implementation Reports*, vol. 17, no. 11, pp. 2265–2300, 2019. doi:10.11124/JBISRIR-2017-003927

[27] A. Forsberg, U. Hammar, A. Ekbom, y R. Hultcrantz, "Post-colonoscopy colorectal cancers in Sweden," *European Journal of Gastroenterology & Hepatology*, vol. 29, no. 7, pp. 855–860, 2017. doi:10.1097/MEG.0000000000000884

[28] L. K. Wanders, S. C. van Doorn, P. Fockens, y E. Dekker, "Quality of colonoscopy and advances in detection of colorectal lesions: a current overview," *Expert Review of Gastroenterology & Hepatology*, vol. 9, no. 4, pp. 417–430, 2015. doi:10.1586/17474124.2015.972940

[29] P. Wieszczy, J. Regula, y M. Kaminski, "Adenoma detection rate and risk of colorectal cancer," *Best Practice & Research Clinical Gastroenterology*, vol. 31, no. 4, pp. 441–446, 2017. doi:10.1016/j.bpg.2017.07.002

[30] Y. Mori, S.-E. Kudo, T. M. Berzin, M. Misawa, y K. Takeda, "Computer-aided diagnosis for colonoscopy," *Endoscopy*, vol. 49, no. 8, pp. 813–819, 2017. doi:10.1055/s-0043-109430

[31] R. Jover *et al.*, "Endoscopist characteristics that influence the quality of colonoscopy," *Endoscopy*, vol. 48, no. 3, pp. 241–247, 2016. doi:10.1055/s-0042-100185

[32] D. Castaneda, V. B. Popov, E. Verheyen, P. Wander, y S. A. Gross, "New technologies improve adenoma detection rate, adenoma miss rate, and polyp detection rate: a systematic review and meta-analysis," *Gastrointestinal Endoscopy*, vol. 88, no. 2, pp. 209–222.e11, 2018. doi:10.1016/j.gie.2018.03.022

[33] F. U. Ertem *et al.*, "Incidence of interval colorectal cancer attributable to an endoscopist in clinical practice," *Gastrointestinal Endoscopy*, vol. 88, no. 4, pp. 705—-711.e1, 2019. doi:10.1016/j.gie.2018.05.012

[34] S. Nazarian, B. Glover, H. Ashrafian, A. Darzi, y J. Teare, "The diagnostic accuracy of artificial intelligence and computer-aided diagnosis for the detection and characterisation of colorectal polyps: A systematic review and meta-analysis. (Preprint)," *Journal of Medical Internet Research*, vol. 23, no. 7, p. e27370, 2021. doi:10.2196/27370

[35] S. Ishaq *et al.*, "Technological advances for improving adenoma detection rates: The changing face of colonoscopy," *Digestive and Liver Disease*, vol. 49, no. 7, pp. 721–727, 2017. doi:10.1016/j.dld.2017.03.030

[36] M. F. Kamiński *et al.*, "Advanced imaging for detection and differentiation of colorectal neoplasia: European Society of Gastrointestinal Endoscopy (ESGE) guideline," *Endoscopy*, vol. 46, no. 5, pp. 435–457, 2014. doi:10.1055/s-0034-1365348

[37] S. Y. Kim, H. S. Kim, y H. J. Park, "Adverse events related to colonoscopy: global trends and future challenges," *World Journal of Gastroenterology*, vol. 25, no. 2, pp. 190–204, 2019. doi:10.3748/wjg.v25.i2.190

[38] A. Rastogi, "Clinical Review: How to Recognize Subtle Lesions in the Colon," *ASGE Leading Edge*, vol. 5, no. 1, pp. 1–11, 2015. [Online]. Disponible en: https://www.asge.org/home/resources/publications/asge-leading-edge/leading-edge-articles/asge-leading-edge-2015-no-1 Último acceso: 23 de julio de 2021.

[39] S. C. van Doorn, R. B. Klanderman, Y. Hazewinkel, P. Fockens, y E. Dekker, "Adenoma detection rate varies greatly during colonoscopy training," *Gastrointestinal Endoscopy*, vol. 82, no. 1, pp. 122–129, 2015. doi:10.1016/j.gie.2014.12.038

[40] S. Muthukuru *et al.*, "Quality of Colonoscopy: A Comparison Between Gastroenterologists and Nongastroenterologists," *Diseases of the colon and rectum*, vol. 63, no. 7, pp. 980–987, 2020. doi:10.1097/DCR.0000000000001659

[41] S. Mansoor, T. Dolkar, y H. El-Fanek, "Polyps and polypoid lesions of the colon," *International Journal of Surgical Pathology*, vol. 21, no. 3, pp. 215–223, 2013. doi:10.1177/1066896913481060

[42] J. E. East *et al.*, "Advanced endoscopic imaging: European Society of Gastrointestinal Endoscopy (ESGE) technology review," *Endoscopy*, vol. 48, no. 11, pp. 1029–1045, 2016. doi:10.1055/s-0042-118087

[43] H. Kato, T. Sakamoto, H. Otsuka, R. Yamada, y K. Watanabe, "Endoscopic diagnosis and treatment for colorectal cancer," en *Colorectal Cancer - From Prevention to Patient Care*, R. Ettarh, Ed., 2012. ISBN 978-953-51-0028-7

[44] C. Lahiff y J. E. East, "Endoscopic approach to polyp recognition," *Frontline Gastroenterology*, vol. 8, no. 2, pp. 98–103, 2017. doi:10.1136/flgastro-2016-100777

[45] Participants in the Paris Workshop, "The Paris endoscopic classification of superficial neoplastic lesions: esophagus, stomach, and colon: November 30 to December 1, 2002," *Gastrointestinal Endoscopy*, vol. 58, no. 6 Suppl, pp. S3–43, 2003. doi:10.1016/s0016-5107(03)02159-x

[46] Endoscopic Classification Review Group, "Update on the Paris classification of superficial neoplastic lesions in the digestive tract," *Endoscopy*, vol. 37, no. 6, pp. 570–578, 2005. doi:10.1055/s-2005-861352

[47] M. S. Ribeiro y M. B. Wallace, "Endoscopic treatment of early cancer of the colon," *Gastroenterology & Hepatology*, vol. 11, no. 7, pp. 445–452, 2015, PMID: 27118940.

[48] R. Lambert *et al.*, "Pragmatic classification of superficial neoplastic colorectal lesions," *Gastrointestinal Endoscopy*, vol. 70, no. 6, pp. 1182–1199, 2009. doi:10.1016/j.gie.2009.09.015

[49] S. C. van Doorn *et al.*, "Polyp morphology: an interobserver evaluation for the Paris classification among international experts." *The American Journal of Gastroenterology*, vol. 110, no. 1, pp. 180–187, 2015. doi:10.1038/ajg.2014.326

[50] L. F. Sánchez-Peralta, L. Bote-Curiel, A. Picón, F. M. Sánchez-Margallo, y J. B. Pagador, "Deep learning to find colorectal polyps in colonoscopy: a systematic literature review," *Artificial Intelligence in Medicine*, vol. 108, p. 101923, 2020. doi:10.1016/j.artmed.2020.101923

[51] T. Utsumi *et al.*, "Polyp detection, characterization, and management using narrow-band imaging with/without magnification," *Clinical Endoscopy*, vol. 48, no. 6, pp. 491–497, 2015. doi:10.5946/ce.2015.48.6.491

[52] F. Wang y A. Preininger, "AI in health: state of the art, challenges, and future directions," *Yearbook of medical informatics*, vol. 28, no. 1, pp. 16–26, 2019. doi:10.1055/s-0039-1677908

[53] Y. LeCun, Y. Bengio, y G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015. doi:10.1038/nature14539

[54] A. Picon, A. Alvarez-Gila, U. Irusta, y J. Echazarra, "Why deep learning performs better than classical machine learning?" *Dyna (Spain)*, vol. 95, no. 2, pp. 119–122, 2020. doi:10.6036/9574

[55] A. Voulodimos, N. Doulamis, A. Doulamis, y E. Protopapadakis, "Deep learning for computer vision: a brief review," *Computational Intelligence and Neuroscience*, vol. 2018, no. Article ID 7068349, pp. 1–13, 2018. doi:10.1155/2018/7068349

[56] D. Shen, G. Wu, y H.-I. Suk, "Deep learning in medical image analysis," *Annual Review of Biomedical Engineering*, vol. 19, no. 1, pp. 221–248, 2017. doi:10.1146/annurev-bioeng-071516-044442

[57] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017. doi:10.1016/j.media.2017.07.005

[58] A. Esteva *et al.*, "Deep learning-enabled medical computer vision," *npj Digital Medicine*, vol. 4, no. 1, p. 5, 2021. doi:10.1038/s41746-020-00376-2

[59] X. Liu *et al.*, "A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis," *The Lancet Digital Health*, vol. 1, no. 6, pp. e271–e297, 2019. doi:10.1016/S2589-7500(19)30123-2

[60] M. Nagendran *et al.*, "Artificial intelligence versus clinicians: Systematic review of design, reporting standards, and claims of deep learning studies in medical imaging," *The BMJ*, vol. 368, pp. 368–m689, 2020. doi:10.1136/bmj.m689

[61] E. Colak, R. Moreland, y M. Ghassemi, "Five principles for the intelligent use of AI in medical imaging," *Intensive Care Medicine*, vol. 47, no. 2, pp. 154–156, 2021. doi:10.1007/s00134-020-06316-8

[62] M. F. Byrne, N. Shahidi, y D. K. Rex, "Will computer-aided detection and diagnosis revolutionize colonoscopy?" *Gastroenterology*, vol. 153, no. 6, pp. 1460–1464.e1, 2017. doi:10.1053/j.gastro.2017.10.026

[63] M. Ashat, J. S. Klair, D. Singh, A. R. Murali, y R. Krishnamoorthi, "Impact of real-time use of artificial intelligence in improving adenoma detection during colonoscopy: a systematic review and meta-analysis," *Endoscopy International Open*, vol. 9, no. 4, pp. E513–E521, 2021. doi:10.1055/a-1341-0457

[64] M. Yamada, Y. Saito, S. Yamada, H. Kondo, y R. Hamamoto, "Detection of flat colorectal neoplasia by artificial intelligence: a systematic review," *Best Practice and Research: Clinical Gastroenterology*, vol. 52–53, p. 101745, 2021. doi:10.1016/j.bpg.2021.101745

[65] Y. Mori *et al.*, "Cost savings in colonoscopy with artificial intelligence-aided polyp diagnosis: an add-on analysis of a clinical trial (with video)," *Gastrointestinal Endoscopy*, vol. 92, no. 4, pp. 905–911.e1, 2020. doi:10.1016/j.gie.2020.03.3759

[66] O. F. Ahmad *et al.*, "Establishing key research questions for the implementation of artificial intelligence in colonoscopy - a modified Delphi method," *Endoscopy*, 2020. doi:10.1055/a-1306-7590 Publicado en línea antes de su edición impresa.

[67] T. J. Sejnowski, "The unreasonable effectiveness of deep learning in artificial intelligence," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 117, no. 48, pp. 30 033–30 038, 2020. doi:10.1073/pnas.1907373117

[68] I. Goodfellow, Y. Bengio, y A. Courville, *Deep Learning*. MIT Press, 2016. [Online]. Disponible en: http://www.deeplearningbook.org/

[69] M. H. Hesamian, W. Jia, X. He, y P. Kennedy, "Deep learning techniques for medical image segmentation: achievements and challenges," *Journal of Digital Imaging*, vol. 32, no. 4, pp. 582–596, 2019. doi:10.1007/s10278-019-00227-x

[70] S. Khan, H. Rahmani, S. A. A. Shah, y M. Bennamoun, *A Guide to Convolutional Neural Networks for Computer Vision*, 2018. ISBN 9781681730219

[71] S. Mohagheghi, A. H. Foruzan, y Y.-W. Chen, "Improving the performance of deep CNNs in medical image segmentation with limited resources," en *Deep Learning in Healthcare: Paradigms and Applications*, Y.-W. Chen y L. C. Jain, Eds. Springer International Publishing, 2020, pp. 79–94. ISBN 978-3-030-32606-7. doi:10.1007/978-3-030-32606-7_5

[72] I. Pacal, D. Karaboga, A. Basturk, B. Akay, y U. Nalbantoglu, "A comprehensive review of deep learning in colon cancer," *Computers in Biology and Medicine*, vol. 126, p. 104003, 2020. doi:10.1016/j.compbiomed.2020.104003

[73] W. Wang *et al.*, "Medical image classification using deep learning," en *Deep Learning in Healthcare: Paradigms and Applications*, Y.-W. Chen y L. C. Jain, Eds. Springer International Publishing, 2020, pp. 33–51. ISBN 978-3-030-32606-7. doi:10.1007/978-3-030-32606-7_3

[74] A. Krizhevsky, I. Sutskever, y G. E. Hinton, "Imagenet classification with deep convolutional neural networks," en *NIPS'12: Proceedings of the 25th International Conference on Neural Information Processing Systems*, 2012, pp. 1097–1105. doi:10.5555/2999134.2999257

[75] K. Simonyan y A. Zisserman, "Very deep convolutional networks for large-scale image recognition," en *The 3rd International Conference on Learning Representations (ICLR2015)*, 2015. [Online]. Disponible en: https://arxiv.org/abs/1409.1556

[76] K. He, X. Zhang, S. Ren, y J. Sun, "Deep residual learning for image recognition," en *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778. doi:10.1109/CVPR.2016.90

[77] G. Huang, Z. Liu, L. van der Maaten, y K. Q. Weinberger, "Densely connected convolutional networks," en *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2261–2269. doi:10.1109/CVPR.2017.243

[78] J. Long, E. Shelhamer, y T. Darrell, "Fully convolutional networks for semantic segmentation," en *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3431–3440. doi:10.1109/CVPR.2015.7298965

[79] O. Ronneberger, P. Fischer, y T. Brox, "U-Net: convolutional networks for biomedical image segmentation," en *Medical Image Computing and Computer-Assisted Intervention. MICCAI 2015. Lecture Notes in Computer Science, vol 9351*, N. Navab, J. Hornegger, W. M. Wells, y A. F. Frangi, Eds. Springer International Publishing, 2015, pp. 234–241. ISBN 978-3-319-24574-4. doi:10.1007/978-3-319-24574-4_28

[80] X. Liu, L. Song, S. Liu, y Y. Zhang, "A review of deep-learning-based medical image segmentation methods," *Sustainability*, vol. 13, no. 3, 2021. doi:10.3390/su13031224

[81] A. Chaurasia y E. Culurciello, "LinkNet: exploiting encoder representations for efficient semantic segmentation," en *2017 IEEE Visual Communications and Image Processing (VCIP)*, 2017, pp. 1–4. doi:10.1109/VCIP.2017.8305148

[82] Z. Eaton-Rosen, F. Bragman, S. Ourselin, y M. J. Cardoso, "Improving data augmentation for medical image segmentation," en *1st Conference on Medical Imaging with Deep Learning (MIDL 2018)*, 2018. [Online]. Disponible en: https://openreview.net/forum?id=rkBBChjiG

[83] J. Schlüter y T. Grill, "Exploring data augmentation for improved singing voice detection with neural networks," en *Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR)*, 2015, pp. 121–126. doi:10.5281/zenodo.1417745

[84] M. Wang y W. Deng, "Deep visual domain adaptation: a survey," *Neurocomputing*, vol. 312, pp. 135–153, 2018. doi:10.1016/j.neucom.2018.05.083

[85] N. Tajbakhsh, L. Jeyaseelan, Q. Li, J. N. Chiang, Z. Wu, y X. Ding, "Embracing imperfect datasets: a review of deep learning solutions for medical image segmentation," *Medical Image Analysis*, vol. 63, p. 101693, 2020. doi:10.1016/j.media.2020.101693

[86] J. Shijie, W. Ping, J. Peiyi, y H. Siping, "Research on data augmentation for image classification based on convolution neural networks," en *2017 Chinese Automation Congress (CAC)*, 2017, pp. 4165–4170. doi:10.1109/CAC.2017.8243510

[87] J. Lemley, S. Bazrafkan, y P. Corcoran, "Smart augmentation learning an optimal data augmentation strategy," *IEEE Access*, vol. 5, pp. 5858–5869, 2017. doi:10.1109/ACCESS.2017.2696121

[88] Y. Shin, H. A. Qadir, L. Aabakken, J. Bergsland, y I. Balasingham, "Automatic colon polyp detection using region based deep CNN and post learning approaches," *IEEE Access*, vol. 6, pp. 40 950–40 962, 2018. doi:10.1109/ACCESS.2018.2856402

[89] K. Cho, J. H. Roh, Y. Kim, y S. Cho, "A performance comparison of loss functions," en *2019 International Conference on Information and Communication Technology Convergence (ICTC)*, 2019, pp. 1146–1151. doi:10.1109/ICTC46691.2019.8939902

[90] G. Keren, S. Sabato, y B. Schuller, "Analysis of loss functions for fast single-class classification," *Knowledge and Information Systems*, vol. 62, no. 1, pp. 337–358, 2020. doi:10.1007/s10115-019-01395-6

[91] V. Ghodrati *et al.*, "MR image reconstruction using deep learning: evaluation of network structure and loss functions," *Quantitative Imaging in Medicine and Surgery*, vol. 9, no. 9, pp. 1516–1527, 2019. doi:10.21037/qims.2019.08.10

[92] B. Kim, M. Han, H. Shim, y J. Baek, "A performance comparison of convolutional neural network-based image denoising methods: the effect of loss functions on low-dose CT images," *Medical Physics*, vol. 46, no. 9, pp. 3906–3923, 2019. doi:10.1002/mp.13713

[93] A. Pathak y R. Maheshwari, "Comparative analysis of different loss functions for deep face recognition," en *ACAI 2019: Proceedings of the 2019 2nd International Conference on Algorithms, Computing and Artificial Intelligence*, 2019, pp. 390–397. doi:10.1145/3377713.3377779

[94] S. Sukhbaatar, J. Bruna, M. Paluri, L. Bourdev, y R. Fergus, "Training convolutional networks with noisy labels," en *3rd International Conference on Learning Representations, ICLR 2015 - Workshop Track Proceedings*, 2015. [Online]. Disponible en: https://arxiv.org/pdf/1406.2080.pdf

[95] T.-Y. Lin, P. Goyal, R. Girshick, K. He, y P. Dollar, "Focal loss for dense object detection," en *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2999–3007. doi:10.1109/ICCV.2017.324

[96] F. Milletari, N. Navab, y S.-A. Ahmadi, "V-Net: fully convolutional neural networks for volumetric medical image segmentation," en *2016 Fourth International Conference on 3D Vision (3DV)*, 2016, pp. 565–571. doi:10.1109/3DV.2016.79

[97] M. A. Rahman y Y. Wang, "Optimizing intersection-over-union in deep neural networks for image segmentation," en *Advances in Visual Computing. ISVC 2016. Lecture Notes in Computer Science, vol 10072*, G. Bebis *et al.*, Eds. Springer International Publishing, 2016, pp. 234–244. ISBN 978-3-319-50835-1. doi:10.1007/978-3-319-50835-1_22

[98] S. S. M. Salehi, D. Erdogmus, y A. Gholipour, "Tversky loss function for image segmentation using 3d fully convolutional deep networks," en *Machine Learning in Medical Imaging. MLMI 2017. Lecture Notes in Computer Science, vol 10541*, Q. Wang, Y. Shi, H.-I. Suk, y K. Suzuki, Eds. Springer International Publishing, 2017, pp. 379–387. ISBN 978-3-319-67389-9. doi:10.1007/978-3-319-67389-9_44

[99] M. Berman, A. R. Triki, y M. B. Blaschko, "The Lovasz-Softmax loss: a tractable surrogate for the optimization of the intersection-over-union measure in neural networks," en *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4413–4421. doi:10.1109/CVPR.2018.00464

[100] N. Abraham y N. M. Khan, "A novel focal Tversky loss function with improved attention U-net for lesion segmentation," en *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, 2019, pp. 683–687. doi:10.1109/ISBI.2019.8759329

[101] H. Kervadec, J. Bouchtiba, C. Desrosiers, É. Granger, J. Dolz, y I. B. Ayed, "Boundary loss for highly unbalanced segmentation," en *Proceedings of The 2nd International Conference on Medical Imaging with Deep Learning*, 2018, pp. 285–296. [Online]. Disponible en: http://proceedings.mlr.press/v102/kervadec19a.html

[102] D. Karimi y S. E. Salcudean, "Reducing the Hausdorff distance in medical image segmentation with convolutional neural networks," *IEEE Transactions on Medical Imaging*, vol. 39, no. 2, pp. 499–513, 2020. doi:10.1109/TMI.2019.2930068

[103] S. Asgari Taghanaki, K. Abhishek, J. P. Cohen, J. Cohen-Adad, y G. Hamarneh, "Deep semantic segmentation of natural and medical images: a review," *Artificial Intelligence Review*, vol. 54, no. 1, pp. 137–178, 2021. doi:10.1007/s10462-020-09854-1

[104] H. Hajiabadi, V. Babaiyan, D. Zabihzadeh, y M. Hajiabadi, "Combination of loss functions for robust breast cancer prediction," *Computers and Electrical Engineering*, vol. 84, p. 106624, 2020. doi:10.1016/j.compeleceng.2020.106624

[105] B. Wang *et al.*, "Deeply supervised 3D FCN with group dilated convolution for automatic MRI prostate segmentation," *Medical Physics*, vol. 46, no. 4, pp. 1707–1718, 2019. doi:10.1002/mp.13416

[106] I. Oksuz *et al.*, "High-quality segmentation of low quality cardiac MR images using k-space artefact correction," en *Proceedings of The 2nd International Conference on Medical Imaging with Deep Learning*, vol. 102, 2019, pp. 380–389. [Online]. Disponible en: http://proceedings.mlr.press/v102/oksuz19a.html

[107] I. Wichakam, T. Panboonyuen, y C. Udomcharoenchaikit, "Real-time polyps segmentation for colonoscopy video frames using compressed fully convolutional network," en *MultiMedia Modeling. MMM 2018. Lecture Notes in Computer Science, vol 10704*, K. Schoeffmann *et al.*, Eds. Springer International Publishing, 2018, pp. 393–404. ISBN 978-3-319-73603-7. doi:10.1007/978-3-319-73603-7_32

[108] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, y J. Liang, "UNet++: a nested U-Net architecture for medical image segmentation," en *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. DLMIA 2018, ML-CDS 2018. Lecture Notes in Computer Science, vol 11045*, D. Stoyanov *et al.*, Eds. Springer International Publishing, 2018. doi:10.1007/978-3-030-00889-5_1

[109] A. K. Mohammed, S. Yildirim-Yayilgan, I. Farup, M. Pedersen, y Ø. Hovde, "Y-Net: A deep Convolutional Neural Network for Polyp Detection," en *29th British Machine Vision Conference*, 2018. [Online]. Disponible en: http://bmvc2018.org/contents/papers/0487.pdf

[110] "EndoVis Grand Challenge," 2021. [Online]. Disponible en: https://endovis.grand-challenge.org/endoscopic_vision_challenge/ Último acceso: 24 de julio de 2021.

[111] J. Bernal *et al.*, "Comparative validation of polyp detection methods in video colonoscopy: results from the MICCAI 2015 Endoscopic Vision Challenge," *IEEE Transactions on Medical Imaging*, vol. 36, no. 6, pp. 1231–1249, 2017. doi:10.1109/TMI.2017.2664042

[112] M. Aziz, R. Fatima, C. Dong, W. Lee-Smith, y A. Nawras, "The impact of deep convolutional neural network-based artificial intelligence on colonoscopy outcomes: A systematic review with meta-analysis," *Journal of Gastroenterology and Hepatology*, vol. 35, pp. 1676–1683, 2020. doi:10.1111/jgh.15070

[113] I. Barua *et al.*, "Artificial intelligence for polyp detection during colonoscopy: a systematic review and meta-analysis," *Endoscopy*, vol. 53, no. 3, pp. 277–284, 2020. doi:10.1055/a-1201-7165

[114] T. K. L. Lui y W. K. Leung, "Is artificial intelligence the final answer to missed polyps in colonoscopy?" *World Journal of Gastroenterology*, vol. 26, no. 35, pp. 5248–5255, 2020. doi:10.3748/wjg.v26.i35.5248

[115] C. Hassan *et al.*, "Performance of artificial intelligence in colonoscopy for adenoma and polyp detection: a systematic review and meta-analysis," *Gastrointestinal Endoscopy*, vol. 93, no. 1, pp. 77–85.e6, 2021. doi:10.1016/j.gie.2020.06.059

[116] C. Hassan, P. Bhandari, G. Antonelli, y A. Repici, "Artificial intelligence for non-polypoid colorectal neoplasms," *Digestive Endoscopy*, vol. 33, no. 2, pp. 285–289, 2021. doi:10.1111/den.13807

[117] T. K. L. Lui, C.-G. Guo, y W. K. Leung, "Accuracy of artificial intelligence on histology prediction and detection of colorectal polyps: a systematic review and meta-analysis," *Gastrointestinal Endoscopy*, vol. 92, no. 1, pp. 11–22.e6, 2020. doi:10.1016/j.gie.2020.02.033

[118] Y. Xu *et al.*, "Comparison of diagnostic performance between convolutional neural networks and human endoscopists for diagnosis of colorectal polyp: a systematic review and meta-analysis," *PLoS ONE*, vol. 16, no. 2, pp. 1–15, 2021. doi:10.1371/journal.pone.0246892

[119] G. Antonelli, M. Badalamenti, C. Hassan, y A. Repici, "Impact of artificial intelligence on colorectal polyp detection," *Best Practice & Research Clinical Gastroenterology*, vol. 52-53, p. 135938, 2020. doi:10.1016/j.bpg.2020.101713

[120] S. Attardo *et al.*, "Artificial intelligence technologies for the detection of colorectal lesions: the future is now," *World Journal of Gastroenterology*, vol. 26, no. 37, pp. 5606–5616, 2020. doi:10.3748/wjg.v26.i37.5606

[121] H. Goyal *et al.*, "Scope of artificial intelligence in screening and diagnosis of colorectal cancer," *Journal of Clinical Medicine*, vol. 9, no. 10, p. 3313, 2020. doi:10.3390/jcm9103313

[122] P. Rasouli *et al.*, "The role of artificial intelligence in colon polyps detection," *Gastroenterology and Hepatology from Bed to Bench*, vol. 13, no. 3, pp. 191–199, 2020. doi:10.22037/ghfbb.v13i3.1866

[123] C. Sánchez-Montes, J. Bernal, A. García-Rodríguez, H. Córdova, y G. Fernández-Esparrach, "Review of computational methods for the detection and classification of polyps in colonoscopy imaging," *Gastroenterology & Hepatology*, vol. 43, no. 4, pp. 222–232, 2020. doi:10.1016/j.gastre.2019.11.003

[124] K.-W. Wang y M. Dong, "Potential applications of artificial intelligence in colorectal polyps and cancer: recent advances and prospects," *World Journal of Gastroenterology*, vol. 9327, no. 34, pp. 5090–5100, 2020. doi:10.3748/wjg.v26.i34.5090

[125] Y. Wang, X. He, H. Nie, J. Zhou, P. Cao, y C. Ou, "Application of artificial intelligence to the diagnosis and therapy of colorectal cancer," *American Journal of Cancer Research*, vol. 10, no. 11, pp. 3575–3598, 2020, PMID: 33294256.

[126] K. O. Kim y E. Y. Kim, "Application of artificial intelligence in the detection and characterization of colorectal neoplasm," *Gut and Liver*, vol. 15, no. 3, pp. 346–353, 2021. doi:10.5009/gnl20186

[127] S. M. Milluzzo, P. Cesaro, L. M. Grazioli, N. Olivari, y C. Spada, "Artificial intelligence in lower gastrointestinal endoscopy: the current status and future perspective," *Clinical Endoscopy*, vol. 54, no. 3, pp. 329–339, 2021. doi:10.5946/ce.2020.082

[128] P. Shen *et al.*, "Real-time use of a computer-aided system for polyp detection during colonoscopy, an ambispective study," *Journal of Digestive Diseases*, vol. 22, no. 5, pp. 256–262, 2021. doi:10.1111/1751-2980.12985

[129] S. Bernard y A. A. Parikesit, "Artificial intelligence in colonoscopy: improving medical diagnostic of colorectal cancer," *Frontiers in Health Informatics*, vol. 9, p. 27, 2020. doi:10.30699/fhi.v9i1.209

[130] N. Hoerter, S. A. Gross, y P. S. Liang, "Artificial intelligence and polyp detection," *Current Treatment Options in Gastroenterology*, vol. 18, pp. 120–136, 2020. doi:10.1007/s11938-020-00274-2

[131] A. Nogueira-Rodríguez *et al.*, "Deep neural networks approaches for detecting and classifying colorectal polyps," *Neurocomputing*, vol. 423, pp. 723–734, 2021. doi:10.1016/j.neucom.2020.02.123

[132] D. Vázquez *et al.*, "A benchmark for endoluminal scene segmentation of colonoscopy images," *Computerized Medical Imaging and Graphics*, vol. 2017, no. 4037190, 2017. doi:10.1155/2017/4037190

[133] D. Jha *et al.*, "Kvasir-SEG: a segmented polyp dataset," en *MultiMedia Modeling. MMM 2020. Lecture Notes in Computer Science, vol 11962*, Y. M. Ro *et al.*, Eds. Springer International Publishing, 2020. ISBN 978-3-030-37734-2. doi:10.1007/978-3-030-37734-2_37

[134] H. Borgli *et al.*, "HyperKvasir: a comprehensive multi-class image and video dataset for gastrointestinal endoscopy," *Scientific Data*, vol. 7, no. 283, 2020. doi:10.1038/s41597-020-00622-y

[135] L. F. Sánchez-Peralta *et al.*, "PICCOLO white-light and narrow-band imaging colonoscopic dataset: a performance comparative of models and datasets," *Applied Sciences*, vol. 10, no. 23, p. 8501, 2020. doi:10.3390/app10238501

[136] S. A. Hick *et al.*, "On evaluation metrics for medical applications of artificial intelligence," *medRxiv*, 2021. doi:10.1101/2021.04.07.21254975

[137] J. F. Ortega-Morán *et al.*, "Medical needs related to the endoscopic technology and colonoscopy for colorectal cancer diagnosis," *BMC Cancer*, vol. 21, p. 467, 2021. doi:10.1186/s12885-021-08190-z

[138] L. F. Sánchez-Peralta, A. Picón, F. M. Sánchez-Margallo, y J. B. Pagador, "Unravelling the effect of data augmentation transformations in polyp segmentation," *International Journal of Computer Assisted Radiology and Surgery*, vol. 15, pp. 1975–1988, 2020. doi:10.1007/s11548-020-02262-4

[139] L. F. Sánchez-Peralta, A. Picón, J. A. Antequera-Barroso, J. F. Ortega-Morán, F. M. Sánchez-Margallo, y J. B. Pagador, "Eigenloss: Combined PCA-Based Loss Function for Polyp Segmentation," *Mathematics*, vol. 8, no. 8, p. 1316, 2020. doi:10.3390/math8081316

[140] K. Wickstrøm, M. Kampffmeyer, y R. Jenssen, "Uncertainty modeling and interpretability in convolutional neural networks for polyp segmentation," en *2018 IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP)*, 2018, pp. 1–6. doi:10.1109/MLSP.2018.8516998

8 Resultados de esta tesis

Contribuciones publicadas

[1] Ortega-Morán, J.F.; Azpeitia, Á.; **Sánchez-Peralta, L.F.**; Bote-Curiel, L.; Pagador, J.B.; Cabezón, V.; Saratxaga, C.L.; Sánchez-Margallo, F.M. Medical needs related to the endoscopic technology and colonoscopy for colorectal cancer diagnosis. BMC Cancer (2021) 21, 467. DOI: 10.1186/s12885-021-08190-z

[2] **Sánchez Peralta L.F.**; Pagador, J.B.; Picón, A.; Calderón, Á.J.; Polo, F.; Andraka, N.; Bilbao, R.; Glover, B.; Saratxaga, C.L.; Sánchez-Margallo, F.M. PICCOLO white-light and narrow-band imaging colonoscopic dataset: a performance comparative of models and datasets. Applied Sciences (2020) 10, 8501. DOI: 10.3390/app10238501

[3] **Sánchez Peralta L.F.**; Picón A.; Sánchez Margallo F.M.; Pagador J.B. Unravelling the effect of data augmentation transformations in polyp segmentation. International Journal of Computer Assisted Surgery and Radiology (2020) 15, 1975-1988. DOI: 10.1007/s11548-020-02262-4

[4] **Sánchez-Peralta, L.F.**; Picón, A.; Antequera-Barroso, J.A.; Ortega-Morán, J.F.; Sánchez-Margallo, F.M.; Pagador, J.B. Eigenloss: combined PCA-based loss function for polyp segmentation. Mathematics (2020) 8, 1316. DOI: 10.3390/math8081316

[5] **Sánchez-Peralta, L.F.**; Bote-Curiel, L.; Picon, A.; Sánchez-Margallo, F.M.; Pagador, J.B. Deep learning to find colorectal polyps in colonoscopy: a systematic literature review. Artificial Intelligence in Medicine (2020) 108, 101923. DOI: 10.1016/j.artmed.2020.101923

Otros artículos publicados

[1] Picon A.; Medela A.; **Sánchez-Peralta L.F.**; Cicchi R.; Bilbao R.; Alfieri D.; Elola A., Glover B.; Saratxaga C.L. Autofluorescence image reconstruction and virtual staining for in-vivo optical biopsying. IEEE Access (2021) 9:32081-32093. DOI: 10.1109/ACCESS.2021.3060926

Capítulos de libro

[1] **Sánchez-Peralta L.F.**, Pagador J.B., Sánchez-Margallo F.M. AIM for colorectal polyps in colonoscopy (2021) en Artificial Intelligence in Medicine. Springer International Publishing. ISBN: 978-3-030-64574-8 – *In press*

Proceedings en Revistas

[1] **Sánchez-Peralta L.F**; Ortega Morán J.F; Saratxaga C.L.; Picón A.; Mündermann L.; Polo F.; Sánchez Margallo F.M. (2021) Towards an unified framework for validating deep learning methods for colorectal polyps: first steps. British Journal of Surgery, 108(S3):iii9s

[2] **Sánchez Peralta, L.F.**, Calderón, A.J., Cabezón, V., Ortega Morán, J.F., Sánchez Margallo, F.M., Polo, F., Saratxaga, C.L., Picón, A. (2019). Systematic acquisition and annotation of clinical cases for the generation of a medical image database. British Journal of Surgery, 106(S2),16

[3] **Sánchez-Peralta, L.F.**, Sánchez-Margallo, F.M., Bote Chacón, J., Soria Gálvez, F., López-Saratxaga, C., Picón Ruiz, A., Pagador, J.B. (2018). Is it necessary to improve the colorectal polyps databases for detection CAD systems based on deep learning? British Journal of Surgery, 105(S2), 5–14.

Comunicaciones a Congresos

[1] **Sánchez-Peralta L.F**; Ortega Morán J.F; Saratxaga C.L.; Picón A.; Mündermann L.; Polo F.; Sánchez Margallo F.M. Hacía un marco unificado para la validación de métodos de deep learning para pólipos colorrectales: primeros pasos. Primer Congreso Virtual de la SEIQ. 11-12 marzo 2021

[2] **Sánchez-Peralta L.F**; Pagador J.B.; Sánchez Margallo F.M.; Bote Curiel L.; Picón A. PCA-based metric reduction for model evaluation in polyp segmentation. 8th European Medical and Biological Engineering Conference (EMBEC 2020) 29 noviembre - 3 diciembre 2020. Portorož, Eslovenia

[3] **Sánchez Peralta L.F.**; Picón A.; Sánchez Margallo F.M.; Pagador J.B. Unravelling the effect of data augmentation transformations in polyp segmentation. 34th International Congress and Exhibition Computer Assisted Surgery and Radiology (CARS2020). 23-27 Junio 2020. Munich

[4] Credi, C., Bocchi, L., Picon, A., Saratxaga, C.L.; **Sánchez Peralta, L.F.**, Ortega Morán, J.F., Bote Chacón, J., Azpeitia Garcia, A., Egia Bizkarralegorra, A., Glover, B., Saverio Pavone, F., Cicchi, R.: Multimodal colon cancer imaging database for disease modeling. 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC'19), Berlin (2019)

[5] Cabezón, V., Egia, A., Belar, O., Calderón, A.J., Polo, F., Etxezarraga, C., Arbide, N., Velasco, J., **Sánchez Peralta, L.F.**, Glover, B., Credi, C., López, C., Picón, A., Bilbao, R.: El Biobanco Vasco como Banco de Imágenes Clínicas y Datos Asociados. Proyecto PICCOLO. IX Congreso Nacional de Biobancos, Oviedo (2018)

[6] **Sánchez-Peralta, L.F.**, Bote-Curiel L., Pagador, J.B, Sánchez-Margallo, F.M., Picón Ruiz, A.: Automatic Segmentation of Colorectal polyps using deep learning. XXXVI Congreso Anual de la Sociedad Española de Ingeniería Biomédica (CASEIB2018), Ciudad Real (2018)

[7] **Sánchez Peralta. L.F.**, Calderón, A.J., Cabezón, V., Ortega Morán, J.F., Sánchez Margallo, F.M., Polo, F., Saratxaga, C.L., Picón, A.: Adquisición y Anotación Sistemática de Casos Clínicos para la Creación de una Base de Datos de Imágenes Médicas. XXIV Congreso de la Sociedad Española de Investigaciones Quirúrgicas, Bilbao (2018)

[8] **Sánchez-Peralta, L.F.**, Sánchez-Margallo, F.M., Bote Chacón, J., Soria Gálvez, F., López Saratxaga, C., Picón Ruiz, A., Pagador, J.B.: ¿Es necesario mejorar las bases de datos de pólipos colorrectales para sistemas CAD de detección basados en deep learning? XXIII Congreso de la Sociedad Española de Investigaciones Quirúrgicas, Madrid (2017)

8.1 Medical needs related to the endoscopic technology and colonoscopy for colorectal cancer diagnosis

Dr. J. Blas Pagador Carrasco y Dr. Artzai Picón Ruiz, como directores de la tesis titulada "Mejoras en la detección, localización y segmentación automática de pólipos colorrectales basadas en técnicas de aprendizaje profundo", certificamos el factor de impacto y la categorización de la siguiente publicación, incluida en la tesis doctoral. Del mismo modo, se especifica la aportación del doctorando.

Autores: Juan Francisco Ortega-Morán, Águeda Azpeitia, **Luisa F. Sánchez-Peralta**, Luis Bote-Curiel, Blas Pagador, Virginia Cabezón, Cristina L. Saratxaga and Francisco M. Sánchez-Margallo

Título: Medical needs related to the endoscopic technology and colonoscopy for colorectal cancer diagnosis

Revista: BMC Cancer

Otros datos: vol 21, pág. 467, año 2021

DOI: 10.1186/s12885-021-08190-z

Factor de impacto (2020): 4.430 (Q2 - Oncology)

Background: The high incidence and mortality rate of colorectal cancer require new technologies to improve its early diagnosis. This study aims at extracting the medical needs related to the endoscopic technology and the colonoscopy procedure currently used for colorectal cancer diagnosis, essential for designing these demanded technologies. Methods: Semi-structured interviews and an online survey were used. Results: Six endoscopists were interviewed and 103 were surveyed, obtaining the demanded needs that can be divided into: a) clinical needs, for better polyp detection and classification (especially flat polyps), location, size, margins and penetration depth; b) computer-aided diagnosis (CAD) system needs, for additional visual information supporting polyp characterization and diagnosis; and c) operational/physical needs, related to limitations of image quality, colon lighting, flexibility of the endoscope tip, and even poor bowel preparation. Conclusions: This study shows some undertaken initiatives to meet the detected medical needs and challenges to be solved. The great potential of advanced optical technologies suggests their use for a better polyp detection and classification since they provide additional functional and structural information than the currently used image enhancement technologies. The inspection of remaining tissue of diminutive polyps (< 5 mm) should be addressed to reduce recurrence rates. Few progresses have been made in estimating the infiltration depth. Detection and classification methods should be combined into one CAD system, providing visual aids over polyps for detection and displaying a Kudo-based diagnosis suggestion to assist the endoscopist on real-time decision making. Estimated size and location of polyps should also be provided. Endoscopes with 360° vision are still a challenge not met by the mechanical and optical systems developed to improve the colon inspection. Patients and healthcare providers should be trained to improve the patient's bowel preparation.

Contribución del doctorando: participated in the design, performance and analysis of the online survey.

Dr. D. José Blas Pagador Carrasco Dr. D. Artzai Picón Ruiz

© The Author(s). 2021 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Ortega-Morán *et al. BMC Cancer* (2021) 21:467
https://doi.org/10.1186/s12885-021-08190-z

BMC Cancer

RESEARCH ARTICLE Open Access

Medical needs related to the endoscopic technology and colonoscopy for colorectal cancer diagnosis

Juan Francisco Ortega-Morán[1*], Águeda Azpeitia[2], Luisa F. Sánchez-Peralta[1], Luis Bote-Curiel[1], Blas Pagador[1], Virginia Cabezón[2], Cristina L. Saratxaga[3] and Francisco M. Sánchez-Margallo[1]

Abstract

Background: The high incidence and mortality rate of colorectal cancer require new technologies to improve its early diagnosis. This study aims at extracting the medical needs related to the endoscopic technology and the colonoscopy procedure currently used for colorectal cancer diagnosis, essential for designing these demanded technologies.

Methods: Semi-structured interviews and an online survey were used.

Results: Six endoscopists were interviewed and 103 were surveyed, obtaining the demanded needs that can be divided into: a) clinical needs, for better polyp detection and classification (especially flat polyps), location, size, margins and penetration depth; b) computer-aided diagnosis (CAD) system needs, for additional visual information supporting polyp characterization and diagnosis; and c) operational/physical needs, related to limitations of image quality, colon lighting, flexibility of the endoscope tip, and even poor bowel preparation.

Conclusions: This study shows some undertaken initiatives to meet the detected medical needs and challenges to be solved. The great potential of advanced optical technologies suggests their use for a better polyp detection and classification since they provide additional functional and structural information than the currently used image enhancement technologies. The inspection of remaining tissue of diminutive polyps (< 5 mm) should be addressed to reduce recurrence rates. Few progresses have been made in estimating the infiltration depth. Detection and classification methods should be combined into one CAD system, providing visual aids over polyps for detection and displaying a Kudo-based diagnosis suggestion to assist the endoscopist on real-time decision making. Estimated size and location of polyps should also be provided. Endoscopes with 360° vision are still a challenge not met by the mechanical and optical systems developed to improve the colon inspection. Patients and healthcare providers should be trained to improve the patient's bowel preparation.

Keywords: Medical needs, Endoscopy, Colorectal cancer, Detection, Classification, CAD software

* Correspondence: jfortega@ccmijesususon.com
[1]Jesús Usón Minimally Invasive Surgery Centre, Ctra. N-521, Km 41.8, 10071 Cáceres, Spain
Full list of author information is available at the end of the article

© The Author(s). 2021 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Background

Colorectal cancer (CRC) is the third most commonly diagnosed cancer globally, representing 10% of all cancer diagnoses. In terms of mortality, CRC is the second most deadly cancer worldwide, causing about 9.4% of deaths [1]. Adenocarcinoma accounts for 96% of all CRCs [2]. Despite this, cure rates can reach up to 90% if patients are diagnosed in early stages of the disease [3]. Consequently, this would reduce the duration and costs of diagnosis and treatment, as well as future reoperations [4].

Screening programmes allow for early detection of CRC. Colonoscopy, which is the main detection method, reveals that up to 40% of patients have one or more polyps [5], with potentially non-malignant hyperplastic polyps being the 29–42% of them, and neoplastic ones with malignant potential the rest [5]. However, almost 30% of the polyps are not detected [6], so it is important to improve the adenoma detection rates (ADR), as every 1% of its increase is associated to a 3% decrease of CRC risk, and 5% decrease of mortality related to CRC [7]. In the current gold standard procedure (Fig. 1), all polyps (both hyperplastic and neoplastic) are resected and sent to histopathological analysis for diagnosis, although strategies as "diagnose and leave" for hyperplastic polyps or "resect and discard" for neoplastic polyps could be followed, depending on the used diagnostic technique and the experience of the endoscopist [8, 9]. This standard diagnostic clinical procedure still depends on biopsy, tissue sample preparation and detailed analysis by an expert pathologist, which includes extraction, preparation, cutting, and staining with Hematoxylin-Eosine to assess the morphological pattern. This protocol implies high diagnostic time and costs and may unnecessarily expose patients to the risks associated to polypectomy, besides the high psychological impact that the waiting time might cause [10]. While adenomatous polyps have malignant potential and must be resected to protect against CRC, hyperplastic polyps have not and can be left.

Recent large multi-centre studies [11, 12] have used advanced imaging modalities such as narrow band imaging (NBI) and visual classification schemes, such as NBI International Colorectal Endoscopic (NICE) or vascular pattern intensity, to differentiate hyperplastic from adenomatous polyps during colonoscopy. However, these procedures are not a reliable enough as replacement method for histological analysis, with accuracy well below the recommended levels, and unsuitable to be applied outside expert academic centres as it highly depends on the clinician's experience and training [13, 14].

On the other hand, clinicians lack tools for the assessment of lesion margins, both prior and after resection. Incomplete polyp resection, especially when larger lesions are removed in piecemeal fashion, significantly increases the tumour recurrence probability [15] and might be the cause of post-colonoscopy CRC in up to 50% of the cases [16, 17]. Regretfully, it is currently difficult to study the progression of partially removed polyps [18].

Moreover, when a polyp invades the submucosa, it is classified as T1 CRC and since the risk of lymph node metastasis increases, surgery is recommended [19]. Although the percentage of lymph node metastasis is lower than 15% [20, 21], determining in-situ the invasion depth could allow identifying patients with higher risk of lymph node metastasis to provide them adequate treatment, as cure rates of T1 CRC patients reaches up to 85% with only endoscopic treatment [19]. However, determining invasion depth is still a challenging problem for non-invasive existing techniques such as endoluminal ultrasound [22] or NBI [23].

Therefore, improved diagnostic techniques are required to differentiate hyperplastic from neoplastic

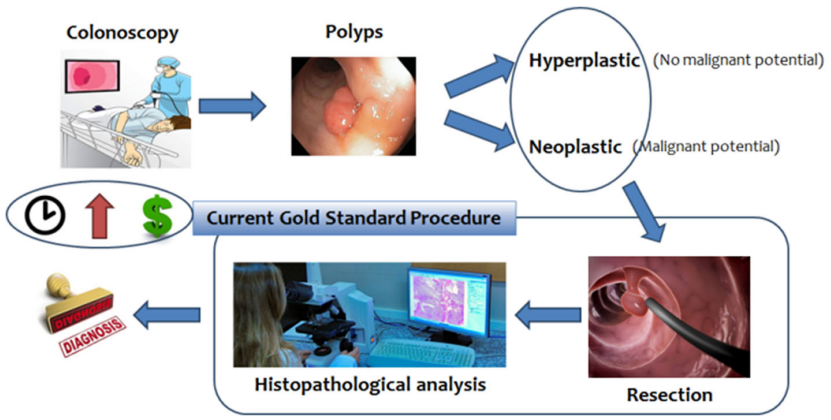


Fig. 1 Current gold standard procedure

polyps, allowing in situ non-invasive assessment, safe characterization and most appropriate resection method of lesions during colonoscopy. New systems that provide accurate and objective optical identification with software support at the time of colonoscopy would reduce time and costs of diagnosis. In fact, the European Society of Gastrointestinal Endoscopy has recently identified the use of optical diagnosis as one of the key research questions that gastroenterology faces [24]. However, in order to improve the design and development of such new technologies, it is important to previously identify precise medical needs and colonoscopy procedural constraints, which is the objective of this work.

Methods

Design and materials

To identify medical needs, a 2-step process has been followed. Initially, a reduced number of endoscopists were interviewed to extract a draft list of needs, which was later confirmed and improved by surveying a broader number of clinicians.

Interviews

The semi-structured interview technique allows respondents to freely express their point of view in their own terms. It has been used in the first step of this study, as it is considered a versatile and flexible method [25]. The interviewer followed an interview guide, which included a list of questions in a predefined, preferred order that should be covered during the conversation. The interview guide included four blocks of questions (see Additional file 1), ranging from generic to concrete, which were based on literature review relevant findings. Such review was conducted with the terms *endoscopy, colonoscopy, interview, colorectal cancer* using PubMed, Google Scholar, Web of Science for studies published between 1990 and 2015. These questions were revised by independent clinicians, including gastroenterologists, endoscopists and pathologists, to assure that the final set of questions covered all relevant issues. Each interview was expected to last approximately 30 min and was carried out in the clinicians' mother tongue. Endoscopists were the selected profile for these interviews according to their knowledge and experience. Both tape-recording and written-notes strategies were used to register these open-end questions interviews. Later, these records were transcribed and further analysed together with written notes.

Survey

An online questionnaire was developed using Survey Monkey [26], aiming to contrast conclusions drawn from the interviews with a broader number of clinicians. A preliminary draft questionnaire was designed based on the most relevant results from the interviews, but also supported by a review of related studies. The search strategy for such review was conducted using PubMed, Google Scholar, Web of Science for studies published between 1990 and 2015, with the terms *endoscopy, colonoscopy, questionnaire, survey, colorectal cancer*. In first place, after choosing the language (Spanish, English or German), the respondent was asked to accept an informed consent to access the questionnaire. It contained three blocks of questions: 1) demography; 2) detection of colorectal lesions; and 3) characterization of colorectal lesions. A survey pilot test was performed with a set of experts to estimate its duration, check the understandability, and detect words that might lead to confusion. The final questionnaire can be found in Additional file 2. The questionnaire was distributed via e-mail to previously identified contacts from participating institutions and through relevant scientific societies. The e-mail body text described the study objective and specified indications to fulfil the questionnaire. All answers were kept anonymous.

Participants

Interviews

Four endoscopists from "Hospital San Pedro de Alcántara" in Cáceres (Spain) and two from the "Hospital de Basurto" in Bilbao (Spain) were interviewed. The theoretical saturation criterion was used to determine the final number of interviews.

Survey

To conduct sampling of endoscopists participating in the study, the non-probability technique called convenience sampling has been selected. In all, 133 endoscopists from 15 countries participated in the survey and 103 (75.74%) completed answers were gathered. Incomplete answers were excluded from the analysis.

Ethical considerations

This study exempted from review by the Institutional Review board because it was not within the scope of Law 14/2007 of 3rd July on Biomedical Research. This national regulation indicates that human health-related research involving invasive procedures needs approval by the ethical committee, but in this study the participants were not involved in invasive procedures. For the same reason, it was not considered necessary for participants to sign a written informed consent. Only verbal informed consent was considered sufficient. All participants were informed about the objectives, length and procedure of the study, the institutions responsible for the research, the contact people, the confidentiality and anonymity of the research data, and that their

participation was voluntary, and they could withdraw from the study at any time if they wished to.

Analysis

Interviews

Topic modelling is a statistical analysis technique widely used in free text analysis that automatically identifies latent topics in a collection of documents, and derives hidden patterns exhibited by this collection. This technique is applied to any subject that contains free-text data, including interviews [27], and also in the health care area [28]. Within the possible implementations, the Latent Dirichlet Allocation (LDA) method [29] has been selected. More deeply, by using LDA we estimate the distribution of topics per document and the distribution of words per topic in a collection of documents. This latter distribution allows representing most relevant words in word cloud charts.

Survey

The questionnaire includes categorical questions, except for two open-ended ones, so they are described using counts and percentages. To compare independents samples, independent χ^2 tests with Fischer correction were used. Considered study groups were: 1) experts (> 1000 colonoscopies performed) vs non experts (≤1000 colonoscopies performed), 2) qualified endoscopist with special interest in cancer screening and/or therapeutic endoscopy vs. qualified endoscopist without any special interest, and 3) experts qualified endoscopists with interest in cancer screening vs experts qualified endoscopists without any special interest. A p value under 0.05 was considered statistically significant. Statistical analysis was performed with IBM SPSS Version 23 statistical package.

Results

Demographics

Interviews

A total of six endoscopists from two Spanish hospitals, "San Pedro de Alcantara" in Cáceres and "Hospital de Basurto" in Bilbao, were interviewed. Five of them were male and only one female, being between 32 and 55 years old. Median age was 51.5 [95% Confidence Interval: 44.75;55] years and all of them were senior doctors with more than a thousand of colonoscopies performed.

Survey

In all, 120 (90.2%) endoscopists, from a total of 133 surveyed, agreed to complete the survey. However, 17 (14.2%) did not respond to any question so 103 questionnaires were finally analysed, 68.0% in Spanish, 25.2% in English and the remaining 6.8% in German. Nearly all surveyed endoscopists (95.2%) were European (namely, from Spain, Germany, Switzerland, Italy, Macedonia, Poland, Portugal, Romania, Slovenia, Sweden, United Kingdom), and only three from South America (Paraguay and Peru), one from Asia (India) and another one from Africa (Egypt). 91.2% of surveyed endoscopists were qualified and over two-thirds (72.1%) were interested in cancer screening or therapeutic endoscopy. Overall, 78.2% of participants had performed more than 1000 colonoscopies during their career.

Global results

After analysing the survey results, no significant differences have been found in either of the three considered study groups. This analysis suggests that existing medical needs would be equally perceived by all endoscopists, and would be not biased by belonging to one group. However, this should be considered with caution as qualified endoscopists with special interest in cancer screening and/or therapeutic endoscopy and expert endoscopists are overrepresented. Therefore, only global results from interviews and surveys of all participants are discussed, from which the medical needs regarding the colonoscopy procedure and the used technology can be extracted and divided in three different fields: clinical needs, Computer Aided Diagnosis (CAD) system needs and operational/physical needs. Verbatim quotes from participants are provided in Additional file 3.

Clinical needs

From a clinical point of view, the main concern for endoscopists is to improve polyp detection, especially for flat polyps, to reduce the polyp missing rate. Once polyps have been detected, the next identified need is to minimize the resection of hyperplastic polyps following the "diagnose and leave" strategy. Furthermore, this concern leads to the clinical need related to in-situ classification of detected polyps, so that either the "resect and discard" or "diagnose and leave" strategies can be followed if deemed necessary. This way, costs associated to extraction, processing and analysis of hyperplastic polyps could be reduced, as well as the time for diagnosis, minimizing negative psychological impact in patients, as stated by an endoscopist. Clinicians also need to reduce recurrence rate by assessing lesion margins and inspecting the remaining tissue, avoiding re-interventions which negatively affect the patient prognosis and increase the risk of complications (such as bleeding, perforation, etc.). Equally important is to assess the depth of submucosal invasion for detected lesions, as it is an essential parameter to decide whether endoscopically resect early colorectal cancerous lesions.

Survey results (Table 1) corroborate these abovementioned needs from the interviews. Endoscopists demanded tools to improve the polyp detection. They

Table 1 Clinical Needs

	N	%
An automated system would be preferred	81	88.0
For which type of lesion would a Computer Aided Diagnosis (CAD) system be MOST useful?		
Flat/Depressed Polyps	87	89.7
Sessile Polyps	3	3.1
Pedunculated Polyps	7	7.2
What information would you like to obtain the MOST?		
Infiltration probability	70	83.3
Depth estimation	60	71.4
Diagnosis suggestions	55	65.5
Margin estimation	53	63.1
Grade	30	35.7

felt that current methods for detection are insufficient and therefore that current approaches could be improved by using advanced technologies for polyp detection and characterisation. The rapid identification and classification of polyps was considered as clinically useful. In fact, 88.0% of respondents indicated that they would like to have an automated system that would suggest the diagnosis without having to make an interpretation of the lesion.

The overall preference was a tool to detect flat or depressed polyps, as 89.7% of endoscopists indicated that this assistance would be more beneficial for this type of lesions, reflecting the consensus that they are more difficult to detect than elevated lesions. The most desired characteristic to be provided by such system is the infiltration probability (83.3%), followed by the estimation of the polyp depth (71.4%), a diagnostic suggestion (65.5%) and polyp margins (63.3%). All these features would assist on deciding how to proceed and potentially avoid unnecessary biopsy or polypectomy if demonstrated to be of adequate accuracy.

Computer aided diagnosis (CAD) system needs

According to the results from the interviews, endoscopists expect the software, or computer aided diagnosis (CAD) systems, to support the polyp detection and diagnosis with additional information for a better diagnosis. They also demand that the CAD system provide them with information about physical characteristics of the polyps, such as the size, and its location. Besides including polyp detection and characterization in the CAD system, it is also important to accurately visualize the polyp contour for safe margin resection and avoiding thus recurrence rates. Clinicians stated that a CAD system that indicates polyp invasiveness or remaining adenomatous tissue left after resection is needed, as it is currently not available. Finally, for a better understanding and a user-friendly environment, all this information provided by the CAD system is requested to be mainly visual.

Survey also shows these trends (Table 2), as most endoscopists prefer that the CAD system provides feedback as visual cues to show polyp detection directly highlighting the polyp on the endoscopic image (also called augmented display) (72.7%) instead of being shown on the screen away from it (43.4%). Regarding polyp diagnosis suggestion (classification), a high proportion of endoscopists (83.3%) preferred a 'traffic light'

Table 2 CAD System Needs

	N	%
According to your experience, which is the MOST helpful indicator to establish the malignancy of a lesion?		
Paris classification	21	23.3
Vascularity	12	12.9
Lesion surface (granularity / no granularity)	11	12.1
Kudo's pit pattern	44	47.8
Existing methods are NOT simple or reproducible enough	31	34.1
How would you prefer to be alerted to the detection of a polyp?		
Audible alarm	11	11.1
Visual cue on screen (away from the endoscopic image)	43	43.4
Highlighting the polyp on the endoscopic image (augmented display)	72	72.7
How would you like to receive the feedback on the diagnosis suggestion?		
Audible alarm	3	3.6
Traffic light cues (Green – benign, Yellow – pre-malign, Red – malign)	70	83.3
Text	26	31.0
I would completely remove large non polypoid lesions with more confidence if I had a CAD system	79	86.8
The CAD system could contribute to detect the residual lesion of a piecemeal polypectomy scar	84	92.3

display, with different colours representing different histological predictions, rather than text (31.0%) or an audible alarm (3.6%).

On the other hand, the most helpful method to establish the malignancy of polyps is the Kudo's pit pattern (47.8%), over the Paris classification (23.3%) or the vascularity (12.9%). Therefore, future developments should consider presenting information in this manner, although more than half of the participants (65.9%) think that the existing methods to establish the malignancy of a lesion are not simple or reproducible enough.

Operational/physical needs

Interviews with panels of expert endoscopists gave also insights into the operational needs. Firstly, the morphology of the colon itself makes it difficult to visualize and detect certain polyps. For that, they demanded better lighting inside the colon to perform the colonoscopy, shortening thus the time spent in reviewing or cleaning dark areas. To meet the need for a better visualization and exploration of the mucosa, the use of flexible/rotatory tips in the endoscope should be further explored. This way, polyps located in hidden areas of the colon could be detected by just rotating the tip. However, the flexibility of the instruments to be inserted is also a concern for endoscopists, as they should be flexible enough to pass through the working channel, but rigid enough to transmit the movements and force from outside. Finally, endoscopists would like to use equipment with better image quality.

LDA results

The interviews to analyse consist of a collection of six documents, one for each interviewed endoscopist. After applying the LDA method, two topics that best describe this collection were found. For each topic, the words distributions were represented as word cloud charts, so the larger the occurrence of a word is, the bigger and bolder it appears in the chart. Details of the LDA results are provided in Additional file 4.

The first topic revolves around the patient (counted 22 times). This might be because some technical problems during colonoscopy are mainly related to the patient. While some can be easily overcome, such as the poor preparation of the colon before the colonoscopy, others are intrinsic to the patient, such as the anatomy of the colon in terms of folds or haustrum. In any case, they complicate to see all polyps and dark zones, as well as the progress of the endoscope. Moreover, since colonoscopy has to be performed in a limited time, these patient-related problems also reduce the exploration time. Lastly, reducing the time until the diagnosis has not much impact on the patient's prognosis but rather an executive/management impact.

In the second topic, *polyp* stands out over the rest of words (counted 36 times), suggesting that the focus is placed on correctly knowing how to detect and identify the polyp based on endoscopy imaging, as supported by other words in the topic, such as the ones equivalent to *polyp,* like *lesion*, *injury,* or *adenomatous*, other related to placements of polyp, like *position*, *area*, or *colon*, or actions about the polyp, like *measure*, *detect*, *locate*, or *identify*.

Discussion

The findings of this study provide information from the endoscopists' point of view about the main needs in the context of endoscopic technology and the colonoscopy procedure currently used for the diagnosis of CRC. Results from interviews and questionnaires revealed that needs might be clustered into three different groups: (a) clinical needs, mainly related to problems in polyp detection and classification, especially flats polyps, as well as their location, size, margins and penetration depth or invasiveness more precisely during colonoscopy; (b) CAD system needs, demanding visual information/alarms for assisting on the polyp characterization and diagnosis, and (c) operational/physical needs, especially in terms of equipment limitations related to image quality and colon lighting, poor bowel preparation or the flexibility of the endoscope tip.

Clinical needs

For better in-situ polyp detection and classification, as demanded by clinicians, various technologies are available to enhance the endoscopic image. Firstly, the NBI highlights the contrast between vascular structures and the surrounding mucosa [30], while the Flexible spectral Imaging Color Enhancement (FICE) technique, and its extension Blue Laser Imaging, consider morphological features, pit pattern and vessel characteristics of polyps [31, 32]. On a different approach, the linked color imaging finds pixel color differences in neoplastic and healthy tissue [33]. Lastly, the iScan provides real time virtual chromoendoscopy for a detailed view of the mucosal and vascular patterns [34].

Besides these techniques, the standard polyp detection with white light colonoscopy can be improved further by means of fluorescence [35, 36] or autofluorescence [31] imaging. This last approach does not use exogenous agents employing both visible [37] and near-infrared [38] light. In contrast to this, endocytoscopy is a high-resolution microendoscopy technology that detects fluorescence after the systemic or topical administration of an exogenous fluorophore [39].

The emerging advanced optical imaging technologies, such as confocal fluorescence microscopy (CFM), optical coherence tomography (OCT), Raman spectroscopy

(RS), hyperspectral spectroscopy or multi-photon tomography (MPT), show great potential for assisting clinicians in the early detection of cancerous diseases [40]. CFM provides real-time microscopic images underneath the surface mucosa of 40–70 μm depth, but requires expensive equipment, significant technical expertise and exogenous fluorophore. In case of OCT, this technology provides high-resolution assessment of subsurface structures without contrast agent. However, there might be a delay in image acquisition and in vivo studies in humans are still lacking. RS details the tissue composition at a microscopic level and spectral information is suitable for automated analysis, but further characterization in humans and the development of endoscopic platforms is required. Finally, MPT high-resolution images show histologic features and changes in tissue without contrast agent, but further characterization in humans and development of endoscopic platforms are missing. Thus, combining imaging technologies that offer complementary information so flaws can be compensated is at stage on the current development of new diagnostic solutions [31].

The above-mentioned enhanced imaging techniques have been commonly applied to identify and characterize colonic lesions but they are less focused on identifying incomplete excision and residual tissue during polypectomy or recurrent disease on follow up, even though there is some evidence of its potential [41]. As for NBI, its effectiveness has not been shown to determine complete excision of diminutive polyps (< 5 mm) [42], but the detection of recurrence when examining previous scars was enhanced when combined with high definition (HD) white light [43]. On the other hand, chromoendoscopy with 0.13% indigo carmine has showed a significant increase in residual polyp detection compared to alone while light endoscopy in predominantly larger polyps (> 5 mm) [44]. Alternatively, the Fourier transform infrared spectroscopy, that provides chemical analysis at the molecular level, helps determining whether the resection margin contains foci with recurrence potential [45].

Finally, regarding the clinical need to estimate the depth of submucosal infiltration of polyps, the endoscopic ultrasonography (EUS) and narrow band imaging-magnifying endoscopy (NBI-ME) are often used. The axial images of the tumors produced by EUS could detect the invasion of the normal layered structure of the colorectal wall [46], while the NBI-ME combination could provide visualization of capillary vessels and their fine structure in the surface layer [47, 48]. A comparative study has shown that NBI-ME is more appropriate for estimating the early CRC invasion depth before treatment to avoid unnecessary surgery, whereas EUS should be applied to early CRC in which the decision to conduct endoscopic resection is difficult [49]. Furthermore, the probe-based confocal laser endo-microscopy is also useful for the differentiation of normal submucosa from carcinoma infiltration, especially when is accompanied by severe fibrosis. This technique provides more sensitivity and accuracy than magnifying chromoendoscopy, but large-scale prospective studies are needed to further evaluate the clinical impact of its use during endoscopy [50].

CAD system needs

The imaging technologies should be complemented with advanced image processing methods that facilitate the detection, analysis and diagnosis of polyps on real time through a CAD system.

On the one hand, many detection methods based on artificial intelligence are already available, although clinical evaluation is still missing [51]. ENDOANGEL helps improving the ADR by identifying endoscope slipping and previously seen frames, thus avoiding blind spots [52]. Similarly, the Automated Polyp Detection Software (APDS) [53] analyses colours, structure, texture and motion information of polyps, and marks a region of interest with small green rings on the endoscopic image. This visualization method matches the preferred option in this study. However, APDS should be further improved before its actual clinical application as it did not detect more polyps than those already detected by clinicians [54]. Wang et al. [55] present an automated polyp detection system which was used during colonoscopy providing as output simultaneous visual and sound alarms. This system increased the ADR, predominantly thanks to the detection of hyperplastic polyps and diminutive adenomas, but there was no difference in the detection of advanced adenoma or sessile serrated lesions [53]. Finally, the GI-Genius system shows a faster reaction time for polyp detection than endoscopists while increasing also the ADR [56]. Nevertheless, there is still room for improvement and methods based on deep learning and/or machine learning are still being developed to explore different approaches, such as the extraction of fractal dimension of wireless capsule endoscopy images used for the identification of abnormal frames [57], or the use of a weakly supervised convolutional neural network that can be trained solely with semantically annotated images, indicating whether they contain anomalies or not [58].

Efforts have also been focused on CAD systems for the classification and diagnosis of colonic polyps. EC-CAD performs an analysis of nuclei and texture from endocytoscopic images of colonic tissue to estimate the probability of non-neoplasm, adenoma and invasive cancer [59] and on the same line, but using endocytoscopy with NBI, a CAD system evaluates cellular, glandular, and

vessel structures of colonic polyps and provide the probability of neoplastic or non-neoplastic diagnosis [60]. Similarly, an artificial intelligence model for real-time characterization of colorectal polyps has been developed based on NBI video frames [61]. This system grades the polyp according to the NICE classification into type 1 – hyperplastic or type 2 – adenomatous and provides the associated probability and a credibility score indicating the confidence in the diagnosis.

The estimation of the polyp size influences the diagnosis treatment, so clinicians demands it to be automatically provided by the CAD system, as the polyp size determination through open forceps is subject to error [62]. The challenge is to recover the spatial information for size estimation of a 3D polyp from a colonoscopic 2D image. So far, little progress has been achieved in this regard. A method has been developed for the polyp size classification from colonoscopic videos into under and over 10 mm [63]. Just like the size, the location of the polyp also influences the risk of CRC [64], so it is also interesting to estimate the location of polyps (in mm from the anus), and not only indicate segment they are in.

Operational/physical needs

Besides addressing the above-mentioned clinical and CAD system needs, new imaging technologies with high resolution could also help to reduce the problems of colon lighting for better polyp detection as requested operational/physical needs. CAD systems are mainly focused on the detection of polyps within the endoscopic field of view [54] so their performance is highly dependent on the proper exploration of the colonic mucosa and polyp exposure, as 20% of the colon surface is never surveyed [65]. To solve this limitation, Stanek et al. [66] have developed a CAD system to identify dark or hidden areas of the colon and alert the endoscopist, so that they can be explored more thoroughly. This software assesses the video frame quality and withdrawal spiral motion of the colonoscope to display a green marker in real-time when a quadrant of the image is inspected. Other approach is focused on the development of a 3D map of the colon by predicting the depth from endoscopic images, so the localization of the lumen is used to assist navigation during colonoscopy [67], thus providing a real-time quantitative measure of the colon inspection.

However, besides lighting of the colon, the poor preparation of patients' bowel prior to colonoscopy highly influences polyp detection. Although new trends of split-dose regimen of bowel preparation has been associated with a better colonic cleansing and adenoma detection, the reticent patient attitude towards this split regimen is still a barrier [68].

To maximize the colon exposure, mechanical add-on devices are also used [54], such as a transparent hood, balloon colonoscopy, or wide-angle colonoscopy [69]. These devices have greatly improved polyp detection rates by expanding the visual field behind the folds. Other accessory device-based systems, such as the Endocuff Vision or EndoRings, manipulate the colon folds using radially extended flexible projections [70, 71]. A cheaper and easy-to-access alternative could be the use of a 3D-printed cap adjustable to standard endoscopes with sideoptics [72], which incorporates two micro-cameras fixed to achieve additional views. Approaches like these could meet the operational/physical limitations of current colonoscopes in terms of flexibility or tip stiffness that makes it difficult to find polyps (especially flat ones) located behind the folds or near the colonic flexures.

Availability of endoscopic systems with better image quality is also requested. Current colonoscopes are mainly based on HD images, but sometimes with this resolution images are still blurred, and it is difficult to distinguish the different tissues. Nowadays, some deep learning algorithms are been implemented for imaging improvement, however they can only process 3–6 frames per second (fps) on 4 K videos and 2fps on 8 K videos, far from real-time requirement (25–30 fps) for endoscopy applications [73].

Challenges

According to the discussion, different challenges still need to be addressed to meet the identified needs. Regarding clinical needs, the assessment of lesion margins and inspection of the remaining tissue of diminutive polyps with less than 5 mm with imaging techniques should be further improved, with the aim of reducing recurrence rates, since in this kind of polyps it can reach up to 11% [45].

In case of CAD system needs, additional efforts are encouraged in the development of CAD systems for the detection of flat polyps, as they are problematic for endoscopists and have a considerable prevalence of almost 25% [74]. Moreover, a CAD system that estimates the polyp size in real time, rather than classifying it into categories, should be developed.

The first challenge related to operational needs would be the development of a tip with 360° vision, either with a rotating tip or with a system that would allow peripheral as well as focused vision. As far as we know, new scope modalities have achieved a vision of 330°, such as the Full Spectrum Endoscopy colonoscope or the Third-Eye Retroscope and Third-Eye Panoramic devices [75]. Additionally, the implementation of an accelerator algorithm for 4 K and 8 K videos appropriate for real-time detection should be further addressed.

Future scopes

The combination of OCT and MPT photonics technologies is also a promising approach that can offer high sensitivity and specificity for diagnosis [76], representing an unprecedented powerful clinical tool to be used for both CRC early diagnosis and follow-up. These images modalities provide microscopic structural and functional information, which are not provided by NBI, FICE or similar imaging modalities.

Beyond the advances in the development of CAD systems for either polyp detection or classification, it would be desirable to incorporate both detection and classification functionalities into one single system to complement the normal workflow of endoscopists [61]. As preferred by endoscopists, visual cues should be overlaid onto the clinical display as augmented reality in case of detection, and with a "traffic light" display with different colours for the different histological classifications, based on Kudo's pit pattern. In any case, the final decision on the diagnosis should be made by the clinician, with the CAD software acting as support tool. The opinion of endoscopists indicating that these classification methods (Kudo, Paris, etc.) are not simple and reproducible enough, together with findings from other studies showing that their interobserver variability should be improved [77], make their diagnostic predictions not fully reliable.

On the other hand, educational initiatives for patients to improve their compliance of bowel preparation are encouraged. But due to the foreseeable lack of patient involvement, such educational initiatives should also be addressed to healthcare providers. There are some independent predictive factors of an inadequate bowel preparation, such as diabetes, psychiatric illness, opioid use, active tobacco use, history of inadequate bowel preparation, and Medicaid coverage [78]. Therefore, if healthcare providers or endoscopists previously know the patients with these characteristics, they could foresee it and reinforce their education for a proper bowel preparation.

Regarding operation needs, it has been shown that 4 K ultra HD monitors reduce operative time and intraoperative blood loss in colorectal laparoscopic surgery, so these technologies with higher quality should be widely incorporated in clinical settings for the normal colonoscopy practices, since currently these 4 K ultra HD technologies are mainly available in specialized, high-volume laparoscopic centers [79]. However, the reduction of operative time achieved with these technologies should be taken with caution, as acting quickly with a shorter withdrawal time is also associated with higher interval CRC rates [80].

Conclusions

Findings of this study provide the scientific community and clinicians with knowledge about the problems and needs that currently exist in the standard procedure of colonoscopy and the endoscopic technology used for CRC detection and diagnosis, specially focused on clinical, CAD software and physical/operational needs. Although some initiatives already address such needs, there are still some challenges to be solved.

For polyp detection and classification, the potential of optical technologies, such as OCT and MPT, should be further exploited to provide structural and functional information of colonic tissue and complement current imaging technologies. Similarly, inspection of the remaining tissue of diminutive polyps (< 5 mm) should be further improved to reduce their recurrence rates. Finally, in terms of clinical needs, further studies are needed to analyse the clinical impact of using EUS or NBI for the estimation of the depth of submucosal infiltration during colonoscopy.

To support such imaging technologies, polyp detection and classification methods based on artificial intelligence should be merged in a single CAD system to complement the regular workflow of endoscopists. Such CAD systems should automatically display visual aids for polyp delimitation, Kudo-based diagnosis as well as other relevant metadata, such as the estimated polyp size in mm, the distance from the anus or depth estimation, with the final aim of real-time classification of neoplastic and hyperplastic polyp for choosing the best approach.

Certain operational/physical limitations make it difficult to detect polyps, so further research to develop a tip with 360° vision is highly desirable alongside with wider use of equipment with better image quality than current HD systems and development of real-time processing for polyp detection. Lastly, thoughtful educational proposals for patients and healthcare providers should be implemented to improve the compliance of bowel preparation prior to colonoscopy.

Abbreviations

CAD: Computer-aided diagnosis; CRC: Colorectal cancer; ADR: Adenoma detection rate; NBI: Narrow band imaging; NICE: NBI international colorectal endoscopic; LDA: Latent dirichlet allocation; FICE: Flexible spectral imaging color enhancement; CFM: Confocal fluorescence microscopy; OCT: Optical coherence tomography; RS: Raman spectroscopy; MPT: Multi-photon tomography; HD: High definition; EUS: Endoscopic ultrasonography; NBI-ME: Narrow band imaging-magnifying endoscopy; APDS: Automated polyp detection software; fps: Frames per second

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12885-021-08190-z.

Additional file 1.
Additional file 2.
Additional file 3.
Additional file 4.

Acknowledgments
This work was partially supported by PICCOLO project. This project has received funding from the European Union's Horizon2020 research and innovation programme under grant agreement No 732111. This information reflects only the institution's view, so the Agencies and the Commission are not responsible for any use that may be made of the information it contains. Authors want to thank the institutions that provided the means for sampling the endoscopists for the study, in particular to Dr. Federico Soria Gálvez, from the Endoscopy Unit, Jesús Usón Minimally Invasive Surgery Centre (Cáceres, Spain), Dr. José Carlos Marín, from Department of Gastroenterology, Hospital 12 de Octubre (Madrid, Spain) and the Spanish Society of Digestive Endoscopy (SEED). In addition, the authors would like to gratefully acknowledge all clinicians who willingly participated in the study, both in the interviews and in the questionnaires.

Authors' contributions
JFOM devised the original concept for the article and was the main architect of the study and the major contributor in writing the manuscript. He participated in the design of the online survey and the interviews. He participated in the performance of the interviews to the endoscopist from the "Hospital San Pedro de Alcántara"; AA carried out the interviews to endoscopist from the "Hospital de Basurto" and their subsequent analysis. She also participated in the analysis of the results of the online survey; LFSP participated in the design, performance and analysis of the online survey; LBC participated in the performance of interviews to the endoscopist from the "Hospital San Pedro de Alcántara" and analysed them with LDA method; BP participated in the original conception of the article and contributed in writing it. He also participated in the design of the study; VC participated in the performance of interviews to the endoscopist from the "Hospital de Basurto" and in the design of the online survey; CLS participated in the design of the study and in the coordination and supervision of task performed; FMSM, as Scientific Director of JUMISC, coordinated and supervised all tasks of this study developed by the team work of the Centre. All authors read and approved the final manuscript.

Funding
This work is part of the PICCOLO project, which has received funding from the European Union's Horizon 2020 research and innovation Programme under grant agreement No. 732111.
GR18199, funded by "Consejería de Economía, Ciencia y Agenda Digital, Junta de Extremadura" and co-funded by European Union (ERDF "A way to make Europe").
The funding bodies did not play any roles in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials
The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate
This study was deemed exempt from review by the Ethical Committee of the Jesús Usón Minimally Invasive Surgery Centre because it was not within the scope of Law 14/2007 of 3rd July on Biomedical Research. This national regulation indicates that human health-related research involving invasive procedures needs approval by the ethical committee, but in this study the participants were not involved in invasive procedures. For the same reason, it was not considered necessary for participants to sign a written informed consent. Only verbal informed consent was considered sufficient. All participants were informed about the objectives, length and procedure of the study, the institutions responsible for the research, the contact people, the confidentiality and anonymity of the research data, and that their participation was voluntary, and they could withdraw from the study at any time if they wished to.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Jesús Usón Minimally Invasive Surgery Centre, Ctra. N-521, Km 41.8, 10071 Cáceres, Spain. [2]Biobanco Vasco, Fundación Vasca de Investigaciones e Innovación Sanitaria (BIOEF), Ronda de Azkue, 1, 48902 Barakaldo, Spain. [3]TECNALIA, Basque Research and Technology Alliance (BRTA), Parque Tecnológico de Bizkaia, C/Geldo. Edificio 700, E-48160 Derio, Bizkaia, Spain.

Received: 13 October 2020 Accepted: 14 April 2021
Published online: 26 April 2021

References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2021. https://doi.org/10.3322/caac.21660.
2. Stewart SL, Wike JM, Kato I, Lewis DR, Michaud F. A population-based study of colorectal cancer histology in the United States, 1998–2001. Cancer. 2006; 107(S5):1128–41. https://doi.org/10.1002/cncr.22010.
3. Atkin WS, Edwards R, Kralj-Hans I, Wooldrage K, Hart AR, Northover JMA, et al. Once-only flexible sigmoidoscopy screening in prevention of colorectal cancer: a multicentre randomised controlled trial. Lancet. 2010; 375(Suppl 9726):1624–33. https://doi.org/10.1016/S0140-6736(10)60551-X.
4. Atia MA, Patel NC, Ratuapli SK, Boroff ES, Crowell MD, Gurudu SR, et al. Nonneoplastic polypectomy during screening colonoscopy: the impact on polyp detection rate, adenoma detection rate, and overall cost. Gastrointest Endosc. 2015;82(Suppl 2):370–5. https://doi.org/10.1016/j.gie.2015.01.016.
5. Rajasekhar PT, Mason J, Wilson A, Close H, Rutter M, Saunders B, et al. OC-024 detect inspect characterise resect and discard 2: are we ready to dispense with histology? Gut. 2015;64(Suppl 1):A13.1–A13. https://doi.org/10.1136/gutjnl-2015-309861.24.
6. Heresbach D, Barrioz T, Lapalus MG, Coumaros D, Bauret P, Potier P, et al. Miss rate for colorectal neoplastic polyps: a prospective multicenter study of back-to-back video colonoscopies. Endoscopy. 2008;40(Suppl 04):284–90. https://doi.org/10.1055/s-2007-995618.
7. Becq A, Chandnani M, Bharadwaj S, et al. Effectiveness of a deep-learning polyp detection system in prospectively collected colonoscopy videos with variable bowel preparation quality. J Clin Gastroenterol. 2020;54(Suppl 6): 554-7.
8. Allen JE, Sharma P. Polyp characterization at colonoscopy: clinical implications. Best Pract Res Clin Gastroenterol. 2017;31(Suppl 4):435–40. https://doi.org/10.1016/j.bpg.2017.07.001.
9. Dayyeh BKA, Thosani N, Konda V, et al. 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. Gastrointest Endosc. 2015;81(Suppl 3):502–e1.
10. Miles A, McClements PL, Steele RJ, Redeker C, Sevdalis N, Wardle J. Perceived diagnostic delay and cancer-related distress: a cross-sectional study of patients with colorectal cancer. Psychooncology. 2017;26(Suppl 1): 29–36. https://doi.org/10.1002/pon.4093.
11. Rees CJ, Rajasekhar PT, Wilson A, et al. Narrow band imaging optical diagnosis of small colorectal polyps in routine clinical practice: the detect inspect characterise resect and DISCARD 2 (DISCARD 2) study. Gut. 2017; 66(Suppl 5):887-95.
12. Ignjatovic A, East JE, Suzuki N, Vance M, Guenther T, Saunders B. Optical diagnosis of small colorectal polyps at routine colonoscopy (detect InSpect ChAracterise resect and DISCARD; DISCARD trial): a prospective cohort study. Lancet Oncol. 2009;10(12):1171–8. https://doi.org/10.1016/S1470-2045(09)70329-8.
13. Sakamoto T, Matsuda T, Nakajima T, Saito Y. Clinicopathological features of colorectal polyps: evaluation of the 'predict, resect and discard' strategies. Color Dis. 2013;15(6):e295–300. https://doi.org/10.1111/codi.12210.
14. Hale MF, Kurien M, Basumani P, Slater R, Sanders DS, Hopper AD. Endoscopy II: PTU-233 in vivo polyp size and histology assessment at colonoscopy: are we ready to resect and discard? A multi-Centre analysis of 1212 polypectomies. Gut. 2012;61:A280–1.
15. Pohl H, Srivastava A, Bensen SP, Anderson P, Rothstein RI, Gordon SR, et al. Incomplete polyp resection during colonoscopy—results of the complete adenoma resection (CARE) study. Gastroenterology. 2013;144(Suppl 1):74–80. https://doi.org/10.1053/j.gastro.2012.09.043.
16. Belderbos TD, Pullens HJ, Leenders M, Schipper ME, Siersema PD, van Oijen MG. Risk of post-colonoscopy colorectal cancer due to incomplete adenoma resection:

a nationwide, population-based cohort study. United European Gastroenterol J. 2017;5(Suppl 3):440–7. https://doi.org/10.1177/2050640616662428.

17. le Clercq CM, Bouwens MW, Rondagh EJ, et al. Postcolonoscopy colorectal cancers are preventable: a population-based study. Gut. 2014;63(Suppl 6): 957–63. https://doi.org/10.1136/gutjnl-2013-304880.
18. Adler J, Toy D, Anderson JC, Robertson DJ, Pohl H. Metachronous Neoplasias Arise in a higher proportion of Colon segments from which large polyps were previously removed, and can be used to estimate incomplete resection of 10–20 mm colorectal polyps. Clin Gastroenterol Hepatol. 2019;17(Suppl 11):2277–84. https://doi.org/10.1016/j.cgh.2019.01.047.
19. Kawachi H, Eishi Y, Ueno H, et al. A three-tier classification system based on the depth of submucosal invasion and budding/sprouting can improve the treatment strategy for T1 colorectal cancer: a retrospective multicenter study. Mod Pathol. 2015;28(Suppl 6):872.
20. Tateishi Y, Nakanishi Y, Taniguchi H, Shimoda T, Umemura S. Pathological prognostic factors predicting lymph node metastasis in submucosal invasive (T1) colorectal carcinoma. Mod Pathol. 2010;23(Suppl 8):1068–72. https://doi.org/10.1038/modpathol.2010.88.
21. Nakadoi K, Tanaka S, Kanao H, Terasaki M, Takata S, Oka S, et al. Management of T1 colorectal carcinoma with special reference to criteria for curative endoscopic resection. J Gastroenterol Hepatol. 2012;27(Suppl 6): 1057–62. https://doi.org/10.1111/j.1440-1746.2011.07041.x.
22. Mukae M, Kobayashi K, Sada M, Yokoyama K, Koizumi W, Saegusa M. Diagnostic performance of EUS for evaluating the invasion depth of early colorectal cancers. Gastrointest Endosc. 2015;81(Suppl 3):682–90. https://doi.org/10.1016/j.gie.2014.10.027.
23. Mitsuyama K, Yoshioka S, Kawano H, Tsuruta O, Torimura T. Mo1723 Prediction of tumor grade and invasion depth of colorectal tumors through the scoring of narrow-band imaging findings. Gastrointest Endosc. 2018; 87(Suppl 6):AB497.
24. Rees CJ, Ngu WS, Regula J, Bisschops R, Saftoiu A, Dekker E, et al. European Society of Gastrointestinal Endoscopy – establishing the key unanswered research questions within gastrointestinal endoscopy. Endoscopy. 2016; 48(10):884–91. https://doi.org/10.1055/s-0042-110398.
25. Kallio H, Pietilä AM, Johnson M, Kangasniemi M. Systematic methodological review: developing a framework for a qualitative semi-structured interview guide. J Adv Nurs. 2016;72(Suppl 12):2954–65. https://doi.org/10.1111/jan.13031.
26. SurveyMonkey. 2020. https://www.surveymonkey.com/. Accessed 16 July 2020.
27. Howes C, Purver M, McCabe R. Investigating Topic Modelling for Therapy Dialogue Analysis. *Proceedings of the IWCS 2013 Workshop on Computational Semantics in Clinical Text*. 2013:7–16.
28. Li Y, Rapkin B, Atkinson TM, Schofield E, Bochner BH. Leveraging latent Dirichlet allocation in processing free-text personal goals among patients undergoing bladder cancer surgery. Qual Life Res. 2019;28(Suppl 6):1441–55. https://doi.org/10.1007/s11136-019-02132-w.
29. Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. J Mach Learn Res. 2003;3:993–1022.
30. McGill SK, Evangelou E, Ioannidis JP, Soetikno RM, Kaltenbach T. Narrow band imaging to differentiate neoplastic and non-neoplastic colorectal polyps in real time: a meta-analysis of diagnostic operating characteristics. Gut. 2013;62(Suppl 12):1704–13. https://doi.org/10.1136/gutjnl-2012-303965.
31. Glover B, Teare J, Patel N. A review of new and emerging techniques for optical diagnosis of colonic polyps. J Clin Gastroenterol. 2019;53(Suppl 7): 495–506. https://doi.org/10.1097/MCG.0000000000001222.
32. Bisschops R, Hassan C, Bhandari P, et al. BASIC (BLI adenoma serrated international classification) classification for colorectal polyp characterization with blue light imaging. Endoscopy. 2018;50(Suppl 03):211–20.
33. Kanzaki H, Takenaka R, Kawahara Y, et al. Linked color imaging (LCI), a novel image-enhanced endoscopy technology, emphasizes the color of early gastric cancer. Endosc Int Open. 2017;5(Suppl 10):E1005–13.
34. Basford PJ, Longcroft-Wheaton G, Higgins B, Bhandari P. High-definition endoscopy with i-scan for evaluation of small colon polyps: the HiSCOPE study. Gastrointest Endosc. 2014;79(Suppl 1):111–8. https://doi.org/10.1016/j.gie.2013.06.013.
35. Mayinger B, Neumann F, Kastner C, Haider T, Schwab D. Hexaminolevulinate-induced fluorescence colonoscopy versus white light endoscopy for diagnosis of neoplastic lesions in the colon. Endoscopy. 2010;42(Suppl 01):28–33. https://doi.org/10.1055/s-0029-1243804.
36. Hirche C, Mohr Z, Kneif S, Doniga S, Murawa D, Strik M, et al. Ultrastaging of colon cancer by sentinel node biopsy using fluorescence navigation with indocyanine green. Int J Color Dis. 2012;27(Suppl 3):319–24. https://doi.org/10.1007/s00384-011-1306-5.
37. Moriichi K, Fujiya M, Sato R, et al. Back-to-back comparison of auto-fluorescence imaging (AFI) versus high resolution white light colonoscopy for adenoma detection. BMC Gastroenterol. 2012;12(Suppl 1):75.
38. Shao X, Zheng W, Huang Z. In vivo diagnosis of colonic precancer and cancer using near-infrared autofluorescence spectroscopy and biochemical modeling. J Biomed Opt. 2011;16(Suppl 6):067005.
39. Muldoon TJ, Pierce MC, Nida DL, Williams MD, Gillenwater A, Richards-Kortum R. Subcellular-resolution molecular imaging within living tissue by fiber microendoscopy. Opt Express. 2011;15(Suppl 25):16413–23.
40. Bergholt MS, Zheng W, Ho KY, Teh M, Yeoh KG, Yan So JB, et al. Fiberoptic confocal raman spectroscopy for real-time in vivo diagnosis of dysplasia in Barrett's esophagus. Gastroenterology. 2014;146(Suppl 1):27–32. https://doi.org/10.1053/j.gastro.2013.11.002.
41. O'Morain N, McNamara D. Complete polypectomy and early detection and management of residual disease to reduce the risk of interval colorectal cancers. Acta Oncol. 2019;58(Suppl1):S4–9.
42. Park SK, Ko BM, Han JP, Hong SJ, Lee MS. A prospective randomized comparative study of cold forceps polypectomy by using narrow-band imaging endoscopy versus cold snare polypectomy in patients with diminutive colorectal polyps. Gastrointest Endosc. 2016;83(Suppl 3):527–32. https://doi.org/10.1016/j.gie.2015.08.053.
43. Desomer L, Tutticci N, Tate DJ, Williams SJ, McLeod D, Bourke MJ. A standardized imaging protocol is accurate in detecting recurrence after EMR. Gastrointest Endosc. 2017;85(Suppl 3):518–26. https://doi.org/10.1016/j.gie.2016.06.031.
44. O'Morain N, Shahin A, Ryan B, McNamara D. Chromoendoscopy post-EMR reduces the rate of incomplete polyp resection. Mo 1690 DDW Abstract #2910972. Gastrointerest Endosc. 2018;87(Suppl 6):AB480.
45. Salman A, Sebbag G, Argov S, Mordechai S, Sahu RK. Early detection of colorectal cancer relapse by infrared spectroscopy in "normal" anastomosis tissue. J Biomed Opt. 2015;20(Suppl 7):075007.
46. Kobayashi K, Kida M, Katsumata T, Yoshizawa S, Yokoyama K, Sada M, et al. Clinical role of endoscopic ultrasonography for the diagnosis of early colorectal cancer and selecting the treatment procedure. Dig Endosc. 2003; 15(Suppl 4):298–305. https://doi.org/10.1046/j.1443-1661.2003.t01-3-00262.x.
47. Hirata I, Nakagawa Y, Ohkubo M, Yahagi N, Yao K. Usefulness of magnifying narrow-band imaging endoscopy for the diagnosis of gastric and colorectal lesions. Digestion. 2012;85(Suppl 2):74–9. https://doi.org/10.1159/000334642.
48. Wada Y, Kudo SE, Kashida H, Ikehara N, Inoue H, Yamamura F, et al. Diagnosis of colorectal lesions with the magnifying narrow-band imaging system. Gastrointest Endosc. 2009;70(Suppl 3):522–31. https://doi.org/10.1016/j.gie.2009.01.040.
49. Chao G, Ye F, Li T, Gong W, Zhang S. Estimation of invasion depth of early colorectal cancer using EUS and NBI-ME: a meta-analysis. Tech Coloproctol. 2019;23(Suppl 9):821–30. https://doi.org/10.1007/s10151-019-02076-w.
50. Kim B, Kim YH, Park SJ, Cheon JH, Kim TI, Kim WH, et al. Probe-based confocal laser endomicroscopy for evaluating the submucosal invasion of colorectal neoplasms. Surg Endosc. 2017;31(Suppl 2):594–601. https://doi.org/10.1007/s00464-016-5003-x.
51. Sánchez-Peralta LF, Bote-Curiel L, Picón A, Sánchez-Margallo FM, Pagador JB. Deep learning to find colorectal polyps in colonoscopy: a systematic literature review. Artif Intell Med. 2020:101923.
52. Gong D, Wu L, Zhang J, Mu G, Shen L, Liu J, et al. Detection of colorectal adenomas with a real-time computer-aided system (ENDOANGEL): a randomised controlled study. Lancet Gastroenterol Hepatol. 2020;5(4):352–61. https://doi.org/10.1016/S2468-1253(19)30413-3.
53. Klare P, Sander C, Prinzen M, Haller B, Nowack S, Abdelhafez M, et al. Automated polyp detection in the colorectum: a prospective study (with videos). Gastrointest Endosc. 2019;89(Suppl 3):576–82. https://doi.org/10.1016/j.gie.2018.09.042.
54. Ahmad OF, Lovat LB. Artificial intelligence for colorectal polyp detection: are we ready for prime time? J Med Artif Intell. 2019;2. https://doi.org/10.21037/jmai.2019.09.02.
55. Wang P, Berzin TM, Brown JRG, et al. Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: a prospective randomised controlled study. Gut. 2019;68(Suppl 10):1813–9. https://doi.org/10.1136/gutjnl-2018-317500.

56. Hassan C, Wallace MB, Sharma P, et al. New artificial intelligence system: first validation study versus experienced endoscopists for colorectal polyp detection. Gut. 2020;69(Suppl 5):799-800.
57. Jain S, Seal A, Ojha A, Krejcar O, Bureš J, Tachecí I, et al. Detection of abnormality in wireless capsule endoscopy images using fractal features. Comput Biol Med. 2020;127:104094. https://doi.org/10.1016/j.compbiomed.2020.104094.
58. Iakovidis DK, Georgakopoulos SV, Vasilakakis M, Koulaouzidis A, Plagianakos VP. Detecting and locating gastrointestinal anomalies using deep learning and iterative cluster unification. IEEE Trans Med Imaging. 2018;37(Suppl 10):2196–210. https://doi.org/10.1109/TMI.2018.2837002.
59. Takeda K, Kudo SE, Mori Y, et al. Accuracy of diagnosing invasive colorectal cancer using computer-aided endocytoscopy. Endoscopy. 2017;49(Suppl 08):798–802.
60. Misawa M, Kudo SE, Mori Y, Takeda K, Maeda Y, Kataoka S, et al. Accuracy of computer-aided diagnosis based on narrow-band imaging endocytoscopy for diagnosing colorectal lesions: comparison with experts. Int J Comput Assist Radiol Surg. 2017;12(Suppl 5):757–66. https://doi.org/10.1007/s11548-017-1542-4.
61. Byrne MF, Chapados N, Soudan F, Oertel C, Linares Pérez M, Kelly R, et al. Real-time differentiation of adenomatous and hyperplastic diminutive colorectal polyps during analysis of unaltered videos of standard colonoscopy using a deep learning model. Gut. 2019;68(Suppl 1):94–100. https://doi.org/10.1136/gutjnl-2017-314547.
62. Rex DK, Rabinovitz R. Variable interpretation of polyp size by using open forceps by experienced colonoscopists. Gastrointest Endosc. 2014;79(Suppl 3):402–7. https://doi.org/10.1016/j.gie.2013.08.030.
63. Itoh H, Roth HR, Lu L, et al. Towards automated colonoscopy diagnosis: binary polyp size estimation via unsupervised depth learning. In: In International conference on medical image computing and computer-assisted intervention. Cham: Springer; 2018. p. 611–9.
64. Blanes-Vidal V, Nadimi ES, Buijs MM, Baatrup G. Capsule endoscopy vs. colonoscopy vs. histopathology in colorectal cancer screening: matched analyses of polyp size, morphology, and location estimates. Int J Colorectal Dis. 2018;33(Suppl 9):1309–12.
65. Varayil JE, Enders F, Tavanapong W, Oh J, Wong J, de Groen PC. Colonoscopy: what endoscopists inspect under optimal conditions. Gastroenterology. 2011;140(Suppl 5):S–718.
66. Stanek SR, Tavanapong W, Wong J, Oh JH, Nawarathna RD, Muthukudage J, et al. SAPPHIRE: a toolkit for building efficient stream programs for medical video analysis. Comput Methods Prog Biomed. 2013;112(Suppl 3):407–21. https://doi.org/10.1016/j.cmpb.2013.07.028.
67. Rau A, Edwards PE, Ahmad OF, et al. Implicit domain adaptation with conditional generative adversarial networks for depth prediction in endoscopy. Int J Comput Assist Radiol Surg. 2019;14(Suppl 7):1167–76. https://doi.org/10.1007/s11548-019-01962-w.
68. Radaelli F, Paggi S, Repici A, Gullotti G, Cesaro P, Rotondano G, et al. Barriers against split-dose bowel preparation for colonoscopy. Gut. 2017;66(Suppl 8):1428–33. https://doi.org/10.1136/gutjnl-2015-311049.
69. Yamamoto H, Shinozaki S, Hayashi Y, et al. Advanced treatment and imaging in colonoscopy: the pocket-creation method for complete resection and linked color imaging for better detection of early neoplastic lesions by colonoscopy. Clin Endosc. 2019;52(Suppl 2):107.
70. Konda V, Chauhan SS, Dayyeh BKA, et al. Endoscopes and devices to improve colon polyp detection. Gastrointest Endosc. 2015;81(Suppl 5):1122–9. https://doi.org/10.1016/j.gie.2014.10.006.
71. Van Doorn SC, Van der Vlugt M, Depla ACTM, et al. Adenoma detection with Endocuff colonoscopy versus conventional colonoscopy: a multicentre randomised controlled trial. Gut. 2017;66(Suppl 3):438–45. https://doi.org/10.1136/gutjnl-2015-310097.
72. Walter BM, Hann A, Frank R, Meining A. A 3D-printed cap with sideoptics for colonoscopy: a randomized ex vivo study. Endoscopy. 2017;49(Suppl 08):808–12.
73. Hu J. Achieve Realtime object detection for 4K and 8K endoscopes with FPGA. 2020.
74. Reinhart K, Bannert C, Dunkler D, et al. Prevalence of flat lesions in a large screening population and their role in colonoscopy quality improvement. Endoscopy. 2013;45(Suppl 05):350–6.
75. Aranda-Hernández J, Hwang J, Kandel G. Seeing better-Evidence based recommendations on optimizing colonoscopy adenoma detection rate. World J Gastroenterol. 2016;22(Suppl 5):1767.
76. Cicchi R, Sturiale A, Nesi G, Kapsokalyvas D, Alemanno G, Tonelli F, et al. Multiphoton morpho-functional imaging of healthy colon mucosa, adenomatous polyp and adenocarcinoma. Biomed Opt Express. 2013;4(Suppl 7):1204–13. https://doi.org/10.1364/BOE.4.001204.
77. Gupta S. Trouble in Paris (classification): polyp morphology is in the eye of the beholder. Am J Gastroenterol. 2015;110(Suppl 1):188–91. https://doi.org/10.1038/ajg.2014.411.
78. Shah SA, Zhou E, Parikh ND. Factors affecting outpatient bowel preparation for colonoscopy. Int J Gastrointest Intervent. 2019;8(Suppl 2):70–3. https://doi.org/10.18528/ijgii180039.
79. Mari GM, Crippa J, Achilli P, et al. 4K ultra HD technology reduces operative time and intraoperative blood loss in colorectal laparoscopic surgery. F1000Research. 2020;9(Suppl 106):106.
80. Lund M, Trads M, Njor SH, Erichsen R, Andersen B. Quality indicators for screening colonoscopy and colonoscopist performance and the subsequent risk of interval colorectal cancer: a systematic review. JBI Database System Rev Implement Rep. 2019;17(Suppl 11):2265–300. https://doi.org/10.11124/JBISRIR-2017-003927.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

BMC

8.2 Deep learning to find colorectal polyps in colonoscopy: a systematic literature review

Dr. J. Blas Pagador Carrasco y Dr. Artzai Picón Ruiz, como directores de la tesis titulada "Mejoras en la detección, localización y segmentación automática de pólipos colorrectales basadas en técnicas de aprendizaje profundo", certificamos el factor de impacto y la categorización de la siguiente publicación, incluida en la tesis doctoral. Del mismo modo, se especifica la aportación del doctorando.

Autores: **Luisa F. Sánchez-Peralta**, Luis Bote-Curiel, Artzai Picón, Francisco M. Sánchez-Margallo, J. Blas Pagador

Título: Deep learning to find colorectal polyps in colonoscopy: A systematic literature review

Revista: Artificial Intelligence in Medicine – Número especial: Deep learning Methods for Medical Applications

Otros datos: vol 108, pág. 101923, año 2020

DOI: 10.1016/j.artmed.2020.101923

Factor de impacto (2020): 5.326 (Q1 – Computer Science, Artificial Intelligence, Q1 – Engineering, Biomedical, Q1 – Medical Informatics)

Abstract: Colorectal cancer has a great incidence rate worldwide, but its early detection significantly increases the survival rate. Colonoscopy is the gold standard procedure for diagnosis and removal of colorectal lesions with potential to evolve into cancer and computer-aided detection systems can help gastroenterologists to increase the adenoma detection rate, one of the main indicators for colonoscopy quality and predictor for colorectal cancer prevention. The recent success of deep learning approaches in computer vision has also reached this field and has boosted the number of proposed methods for polyp detection, localization and segmentation. Through a systematic search, 35 works have been retrieved. The current systematic review provides an analysis of these methods, stating advantages and disadvantages for the different categories used; comments seven publicly available datasets of colonoscopy images; analyses the metrics used for reporting and identifies future challenges and recommendations. Convolutional neural networks are the most used architecture together with an important presence of data augmentation strategies, mainly based on image transformations and the use of patches. End-to-end methods are preferred over hybrid methods, with a rising tendency. As for detection and localization tasks, the most used metric for reporting is the recall, while Intersection over Union is highly used in segmentation. One of the major concerns is the difficulty for a fair comparison and reproducibility of methods. Even despite the organization of challenges, there is still a need for a common validation framework based on a large, annotated and publicly available database, which also includes the most convenient metrics to report results. Finally, it is also important to highlight that efforts should be focused in the future on proving the clinical value of the deep learning based methods, by increasing the adenoma detection rate.

Contribución del doctorando: literature search, screening, data extraction, preparation of graphics, tables, and manuscript.

Dr. D. José Blas Pagador Carrasco Dr. D. Artzai Picón Ruiz

0933-3657/ © 2020 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/BY-NC-ND/4.0/).

Artificial Intelligence In Medicine 108 (2020) 101923

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed

Deep learning to find colorectal polyps in colonoscopy: A systematic literature review

Luisa F. Sánchez-Peralta[a,*], Luis Bote-Curiel[a], Artzai Picón[b], Francisco M. Sánchez-Margallo[a], J. Blas Pagador[a]

[a] Jesús Usón Minimally Invasive Surgery Centre, Ctra. N-521, km 41.8, 10071 Cáceres, Spain
[b] Tecnalia, Parque Científico y Tecnológico de Bizkaia, C/ Astondo bidea, Edificio 700, 48160 Derio, Spain

ARTICLE INFO

Keywords:
Colorectal cancer
Deep learning
Detection
Localization
Segmentation

ABSTRACT

Colorectal cancer has a great incidence rate worldwide, but its early detection significantly increases the survival rate. Colonoscopy is the gold standard procedure for diagnosis and removal of colorectal lesions with potential to evolve into cancer and computer-aided detection systems can help gastroenterologists to increase the adenoma detection rate, one of the main indicators for colonoscopy quality and predictor for colorectal cancer prevention. The recent success of deep learning approaches in computer vision has also reached this field and has boosted the number of proposed methods for polyp detection, localization and segmentation. Through a systematic search, 35 works have been retrieved. The current systematic review provides an analysis of these methods, stating advantages and disadvantages for the different categories used; comments seven publicly available datasets of colonoscopy images; analyses the metrics used for reporting and identifies future challenges and recommendations. Convolutional neural networks are the most used architecture together with an important presence of data augmentation strategies, mainly based on image transformations and the use of patches. End-to-end methods are preferred over hybrid methods, with a rising tendency. As for detection and localization tasks, the most used metric for reporting is the recall, while Intersection over Union is highly used in segmentation. One of the major concerns is the difficulty for a fair comparison and reproducibility of methods. Even despite the organization of challenges, there is still a need for a common validation framework based on a large, annotated and publicly available database, which also includes the most convenient metrics to report results. Finally, it is also important to highlight that efforts should be focused in the future on proving the clinical value of the deep learning based methods, by increasing the adenoma detection rate.

1. Introduction

Colorectal cancer (CRC) is defined as a carcinoma, usually an adenocarcinoma, in the colon or rectum. Colorectal cancer is considered primarily as a “lifestyle” disease; its incidence is higher in countries with a diet high in calories and animal fat and with a largely sedentary population [1]. CRC accounts for a 10% of overall new cancer cases worldwide, with a higher incidence rate in developed countries [2]. Only in the United States, it has increased from over 132,000 estimated new cases and nearly 50,000 estimated deaths in 2015 [3] to over 145,000 estimated new cases and 51,000 estimated deaths in 2019 [4]. In Europe, CRC represents the second most common cancer and also the second cause of death from cancer [5].

Nevertheless, an early detection of the CRC increases the 5-year survival rate from 18% when CRC is detected in the highest grade to 88.5% when it is detected in an initial grade due to symptoms. Furthermore, screening programs achieve a significant increase of the survival rate as they allow to start treatment even before the appearance of those symptoms, so up to 222 deaths out of 1000 patients detected with symptomatic CRC could be avoided [6].

Abbreviations: ADR, adenoma detection rate; CAD, computer aided detection; CDDN, cascaded deep decision network; CI, confidence interval; CNN, convolutional neural network; CRC, colorectal cancer; DWD, distance-weighted discrimination; FCN, fully convolutional network; GAN, generative adversarial network; IoU, Intersection over Union; LSTM, long short term memory; mAP, mean average precision; PIVI, preservation and incorporation of valuable endoscopic innovations; RNN, recurrent neural network; ReLU, rectified linear unit; RPN, region proposal network; SVM, support vector machine

* Corresponding author.
E-mail addresses: lfsanchez@ccmijesususon.com (L.F. Sánchez-Peralta), lbote@ccmijesususon.com (L. Bote-Curiel), artzai.picon@tecnalia.com (A. Picón), msanchez@ccmijesususon.com (F.M. Sánchez-Margallo), jbpagador@ccmijesususon.com (J.B. Pagador).

https://doi.org/10.1016/j.artmed.2020.101923
Received 27 August 2019; Received in revised form 3 March 2020; Accepted 1 July 2020
0933-3657/ © 2020 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/BY-NC-ND/4.0/).

Colonoscopy is a standard technique for visual exploration of the colon and rectum by inserting a flexible endoscope through the patient anus [7] and is considered the gold standard for detection and removal of colorectal lesions, associated with important reduction of CRC mortality [8]. The European Society of Gastrointestinal Endoscopy recommends the use of high definition white light endoscopes for detection of colorectal neoplasms in middle risk population [9], as it is estimated that 70–80% of CRC has a sporadic origin [10]. Therefore, it is clear the need for increasing the adenoma detection rate (ADR), defined as the proportion of patients with at least one colorectal adenoma detected among all patients examined by the gastroenterologist, which is both a colonoscopy quality measure and a validated predictor for CRC prevention, having an inverse relationship [11].

Recently, the success of deep learning [12] has also boosted the applications on medical imaging analysis [13], achieving expert performance in several cases [14–16]. Deep learning approaches rely on the ability of networks with several layers to automatically learn hierarchical features characterizing the input data through the application of non-linear operations together with backpropagation for training. Deep learning architectures stack blocks of different types of layers (fully connected, convolutional, pooling or activation layers) to simultaneously be sensitive to minute details and insensitive to large irrelevant details.

Computer-aided detection (CAD) systems have the potential to revolutionize the endoscopic practice by (1) improving the adequacy of inspection technique; (2) providing automatic detection of precursor lesions of CRC; and (3) facilitating real-time diagnosis with optical biopsy [17]. In this review, we focus on the second topic. Traditionally, CAD systems for polyp detection have been based on the manual extraction of polyp features, or so called hand-crafted methods: shape-based [18], texture-based [19,20], depth of valleys-based [21,22] or combined-based [23] methods. Nevertheless, results of the MICCAI 2015 Automatic Polyp Detection in Colonoscopy Videos challenge proved that convolutional neural networks (CNNs) are the state-of-the-art regarding polyp detection methods [24].

Within the scope of this systematic review, three main tasks are considered:

1. Detection: identifying whether a polyp is shown or not in the frame, but information on the polyp location is not given.
2. Localization: identifying the position of the polyp within a given frame, but exact shape of the polyp is not relevant.
3. Segmentation: marking the exact polyp area in a given frame.

Polyp classification (benign vs malign, Paris classification, NICE classification, etc.) is out of the scope of the current systematic review. Besides, classification in this case is done once the presence of the polyp is confirmed and in this review we place the focus on the prior stage (identifying the presence of the polyp in the frame). Therefore, we analyze the published methods for detection, localization and segmentation of colorectal polyps based on deep learning approaches. In this sense, methods are classified according to their main aim, the used database, their approach and the reported metrics. The state-of-art approaches are compared, showing advantages and disadvantages of the different categories, to identify the most auspicious trends.

There are several surveys on deep learning for medical imaging analysis [13,25–31], where at most methods for colorectal polyps identification are roughly analyzed. A more specific review was done by Prasath [32] but the focus was placed on video capsule endoscopy rather than on colonoscopy images. More recently, Ahmad et al. [33] summarized the evidence for clinical applications of computer-aided diagnosis and artificial intelligence in colonoscopy of key studies in a narrative manner. Therefore, to the authors' knowledge, there is no previous systematic review on the proposed topic with a comparative analysis of retrieved works.

Nevertheless, the lack of a common framework makes difficult to provide a reliable comparison of the state-of-art methodologies. The main limitation comes from the different datasets and/or testing set used for reporting results as well as selected metrics. Ideally, all methods should be applied on a common database [34]. Regretfully, this ideal situation is not real in many of the published works, where authors use different datasets, in a different manner, and reporting on different test sets with different metrics. Therefore, in the current review, methods are compared accordingly to their reported results, identifying characteristics that might influence on them. To overcome this heterogeneous situation, efforts have been lately performed by the organization of challenges under the hosting of international congresses in order to unify criteria. It is of special relevance the Endoscopic Vision Challenge [35], which has organized a subchallenge focused on detection, localization and segmentation of polyps in 2015 [24], 2017 and 2018. In these cases, methods are easily and reliably compared. Besides, Vázquez et al. [36] also propose a benchmark for endoluminal scene segmentation of colonoscopy images, with the aim of boostering comparative research.

The main contributions of this systematic review are:

1. We analyze publicly available datasets of colonoscopy images.
2. We provide a comprehensive analysis of polyp detection, localization and segmentation methods based on deep learning, discussing advantages and disadvantages of the different categories.
3. We analyze and discuss the reporting metrics used.
4. We identify future challenges and recommendations based on the findings of the review to be addressed by the scientific community to advance the field.

This review is organized as follows. In Section 2 we present the material and methods to carry out the systematic literature review, including search strategy, study selection and data extraction and management. In Section 3 we show the results of the search and summarize the works found. Then, datasets of colonoscopy images and how they are used are described in Section 4. In Section 5 and its corresponding subsections, we present and discuss the methods for detection (Section 5.1), localization (Section 5.2) and segmentation (Section 5.3) of colorectal polyps using deep learning approaches. We conclude this section with the advantages and disadvantages of each category (Section 5.4). After this, metrics are analyzed in Section 6, to conclude the paper with an analysis from a clinical perspective (Section 7), future challenges and recommendations (Section 8) and final conclusions (Section 9).

2. Material and methods

2.1. Search strategy

The Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) [37] has been followed to perform this systematic review. The basic search string was ("colon" OR "colorectal") AND ("cancer" OR "polyp") AND ("deep learning") AND ("detection" OR "localization" OR "segmentation") and searches were performed on February 2nd 2019 using ACM Digital Library, IEEE Digital Library, Web of Science, PubMed, Science@Direct, Scopus and Springer Link databases. The search string syntax was adapted when necessary, depending on the database requirements. Search was performed on title, abstract and keywords. Previously identified articles were also included in the process.

2.2. Study selection

Studies included in the analysis were full text articles or full-length proceedings published in English. The exclusion criteria were papers published before 2015; in other languages; about a different topic; applying deep learning in a field other than CRC; pure clinical studies;

use of endoscopic capsules as imaging source; not using white light imaging; not using deep learning techniques; short proceedings; and meta-analyses or reviews. Retrieved abstracts were read by two authors (LBC and LFSP), searching for the full text when information in the abstract was not enough to determine its inclusion or exclusion. Full texts of selected abstracts were retrieved (LFSP) and independently revised (AP, LBC, JBP and LFSP) for final agreement on inclusion criteria. The reference list of selected works was also scrutinized for potential interesting articles (LFSP).

2.3. Data extraction

Information from papers selected for analysis was extracted for comparison (JBP, AP and LFSP). Papers were initially categorized according to their main objective (detection, localization or segmentation). For each work, we also identified findings related to the used dataset, the data augmentation approach, the proposed method, the reporting metrics and reproducibility aspects (Table 1).

2.4. Review and data management

Parsif.al[1] was used for abstract review and data management. Search results from each database were imported into the platform. Authors, title, year and journal were automatically extracted, facilitating the selection procedure and the creation of the PRISMA flow diagram. A custom data extraction form was developed in the platform and used for the data extraction process.

3. Results of the systematic search

In all, 1,332 abstracts were found (Table 2 and Fig. 1). 35 papers were previously identified and used to select the search string based on their keywords, while 1,297 were retrieved from searches in the different databases. After removing duplicates, 1,123 abstracts were screened. 1,071 were excluded based on the exclusion criteria and 52 were revised in full text. From those, 33 were included in this review analysis. From manual inspection of references lists, 2 more works were added.

During the revision process, detection was also considered as classification between healthy tissue and polyp classes or polyp and non-polyp classes, but we excluded papers aiming at multi-class classification (polyp among other classes) (such as Pogorelov et al. [38] or Park and Sargent [39]) or polyp classification (neoplastic/non-neoplastic) (such as Ribeiro et al. [40]).

Full text documents were carefully read, and data included in Table 1 were extracted for each work when available in the document. Authors were contacted when deemed necessary for clarification. Table 3 summarizes the 35 articles included in the analysis, together with some relevant aspects. Some works address more than one task, so they appear once per task in the table, resulting in 39 cases. There are 5 book sections (usually proceedings of major conferences), 18 conference proceedings, 11 journal articles and one preprint. To unify criteria for results reporting in Table 1, metrics reported as percentages by authors have been expressed in the normalized range [0, 1].

As it might be expected, there is an overall increasing trend since 2016 in the application of deep learning techniques for polyp detection, localization and segmentation (Fig. 2).

Fig. 3 summarizes the networks used by the different authors. Architectures have been grouped into 4 clusters: (1) CNNs, such as AlexNet [75], VGG16 and VGG19 [76] or GoogLeNet [77], including also residual networks such as ResNet50 [78]; (2) fully convolutional networks (FCNs), based on any CNN architecture, including also the encoder-decoder architectures, such as SegNet [79] or U-Net [80]; (3) generative adversarial networks (GANs), and (4) recurrent neural networks (RNNs), including long short term memory (LSTM). It can be clearly seen that the use of CNNs surpasses the rest of networks in the three analyzed tasks.

A baseline comparison is needed to assess whether the proposed method actually improves the results of the state-of-art. Out of the 35 analyzed works, only 10 of them present a baseline against which the proposed method is compared. The baseline is usually the network on which the proposed method is based, or hand-crafted methods. In other 12 cases, authors provide a comparison of their method against other similar works. The most repeated comparison is against the methods participating in the MICCAI 2015 Automatic Polyp Detection in Colonoscopy Videos challenge [24]. The remaining 13 works only present their work, without any type of comparison.

As for reproducibility, we considered two aspects: the use of public datasets and the availability of the code. Most works (29) use only public datasets, while 3 use only private datasets and other 3 use both public and private ones. The use of proprietary datasets hampers the reproducibility and fair comparison of methods. The code of only three works have been found [36,70,72]. Therefore, Vázquez et al. [36] and Wickstrøm et al. [70] outstand in terms of reproducibility, as the code is available, and they use CVC-EndoSceneStill, which, as it will be explained in the following section, provides a division into training, validation and test sets.

4. Datasets of colonoscopy images

4.1. Currently available public datasets

The creation of large, annotated datasets has also contributed to the tremendous growth of deep learning for the last years. Although they are easily accessible for natural images with different ground truths (i.e. ImageNet [81,82], MSCoco [83,84] or Pascal VOC [85,86]), the limited size of medical imaging datasets is a well-known problem, especially for supervised learning [26]. This situation also applies to colonoscopy images datasets. The type of ground truth highly influences the size of the dataset, since manual annotation of frames is a cumbersome, time-consuming task [87]. Table 4 shows a summary of the currently publicly available datasets of colonoscopy images for polyp detection, localization and segmentation, although prior registration might be required by the dataset owner to grant access to the content. All datasets are mentioned in at least one of the analyzed papers. Although these datasets are widely used in the retrieved papers in this systematic review, some authors also use proprietary datasetss, which compromises a fair comparison of methods and raise concerns about reproducibility, as mentioned in the previous section. Datasets of natural images have been also included in Table 4 for scale comparison between computer vision and biomedical image datasets. Works retrieved in this review only employ these larger datasets to initialize the weights of the network before training.

The organization of challenges under the umbrella of major conferences has meant a great step towards the establishment of a common framework. As a result, most of the currently used datasets were provided within a challenge. The Gastrointestinal Image ANAlysis (GIANA) sub-challenge [91], as part of the EndoVis challenge [35], was last hosted at MICCAI2017 and MICCAI2018. CVC-VideoClinicDB [87,89] was provided as training dataset for the polyp detection task, while CVC-ColonDB, CVC-ClinicDB and CVC-ClinicHDSegment datasets [21,22,36] were provided for the polyp segmentation task. During MICCAI 2015 Automatic Polyp Detection in Colonoscopy Videos challenge [24], two more datasets were released. On one side, ETIS-LARIB was provided as test set for detection on still frames. On the other hand, the ASU-Mayo Clinic [88] was intended for exploring detection on videos. In all cases, the provided ground truth is a binary mask indicating the polyp area. This segmentation is a precise manual delineation except for the CVC-VideoClinicDB dataset, where the polyp area has been

[1] https://parsif.al/

Table 1
Data extraction.

Category	Item	Description
General information	Type of publication	Journal article, conference proceeding, book section (usually proceedings of a major conference), preprint
	Published in	Title of the journal, conference or preprint repository
	Country	Country of the first author affiliation
Objective	Task	Aim of the work: detection, localization and/or segmentation
Data information	Dataset	Name of the public dataset and reference (if available) or proprietary
	Training set	Description of the samples used for training, indicating the number (with and without data augmentation, if applied)
	Validation set	Description of the samples used for validation, indicating the number
	Test set	Description of the samples used for testing, indicating the number
Data augmentation	Approach	Data augmentation approach to generate new training samples: creation of images, on the flow, patch-based or none
	Transformations	If data augmentation is applied, description of the transformations and ranges used
Model	Approach	Type of approach used: feature extractor, classification, patch-based, bounding-box or semantic segmentation
	Architecture	Type of the network (AlexNet, VGG, fully convolutional network, GAN, etc.)
	Loss function	Loss function used for optimization
	Training	Use of pre-trained models or training from scratch
Reporting	Reporting test	Samples used for reporting
	Metrics	Metrics used for reporting (accuracy, Dice, Intersection over Union, etc.)
	Baseline	Whether the authors establish an initial baseline model or compare their results to other works
Reproducibility	Dataset	Whether the dataset is proprietary or publicly available
	Code	Whether the code is publicly available or not

Table 2
Number of abstracts retrieved.

Database	# abstracts
ACM Digital Library	2
IEEE Digital Library	17
Web of Science	43
PubMed	14
Science@Direct	208
Scopus	772
Springer	241
Previously identified	35

approximated to the most convenient elliptical shape. CVC-EndoSceneStill [36] has not been used in any challenge, but it is publicly available. It compiles CVC-ColonDB [22] and CVC-ClinicDB [21] adding ground truth masks for other classes, establishing the distribution of images into training, validation and test sets and indicating the metrics for reporting. Lastly, and not used as much as the previously reported datasets, Kvasir [90] is a multi-class image dataset containing the polyp class among other labels for anatomical landmarks and pathological findings.

There are as well other public datasets used by authors in the current review. The Nerthus database [92], used by Pogorelov et al. [54],

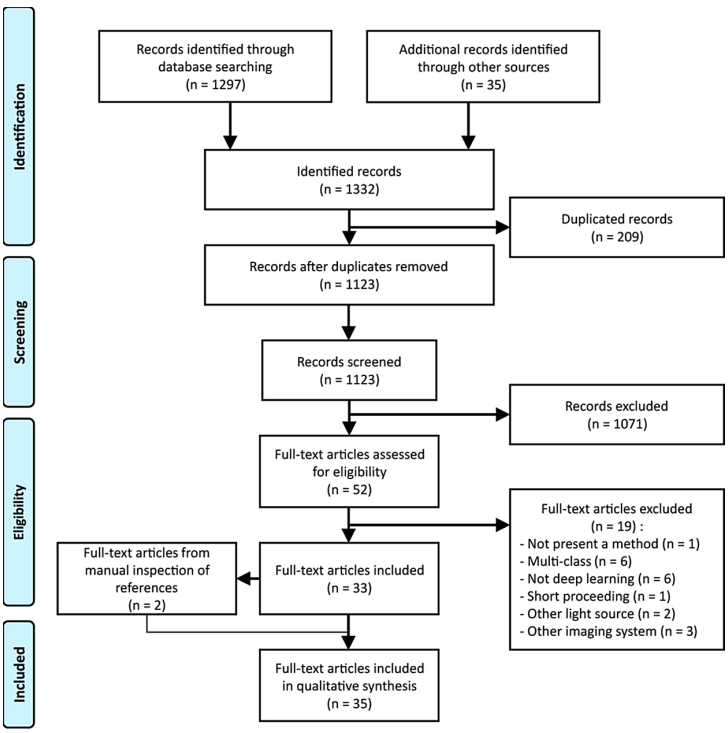


Fig. 1. Literature flow diagram.

Table 3
Summary of reviewed works.

Objective	Type	Approach	Authors	Datasets	Data augmentation	Network	Main results	Baseline	Code
Detection	Hybrid	Feature extractor, patch-based	Taha et al. [41]	CVC-ColonDB	Patches	AlexNet	Prec = 0.927; Rec = 0.960	Comparison	NF
		Feature extractor, bounding-box	Shin et al. [42]	CVC-ClinicDB, ETIS-LARIB, ASU-Mayo	Patches	CNN	Acc = 0.9126; Prec = 0.9271; Rec = 0.9082; Spec = 0.9176	Yes	NF
		Patch-based	Tajbakhsh et al. [43]	ASU-Mayo	Patches	AlexNet	Sens = 0.7 @ FPPF = 0.02	Yes	NF
			Yuan et al. [44]	ASU-Mayo	Patches	AlexNet	Acc = 0.9147; Rec = 0.9176	No	NF
			Tajbakhsh et al. [45]	ASU-Mayo	Patches	AlexNet	Sens = 0.7 @ FPPF = 0.02	Yes	NF
			Tajbakhsh et al. [46]	ASU-Mayo	Patches	CNN	Sens = 0.5 @ FPPF = 0.002	Yes	NF
			Tajbakhsh et al. [47]	ASU-Mayo	Patches	CNN	Sens = 0.5 @ FPPF = 0.002	No	NF
		Classification, patch-based	Axyonov et al. [48]	CVC-ColonDB, ASU-Mayo	N/A	AlexNet	AUC = 0.92; Sens = 0.75 @ FPPF = 0.1	Comparison	NF
	End-to-end	Classification	Akbari et al. [49]	ASU-Mayo	N/A	CNN	Acc = 0.9028; Rec = 0.6832; Prec = 0.7434; Spec = 0.9497; FPPF = 0.06	Comparison	NF
			Aksenov et al. [50]	CVC-ColonDB, ASU-Mayo	Creation	CNN	TPR ≈ 0.98 @FPR = 0.025	No	NF
			Itoh et al. [51]	Proprietary	N/A	C3dNet	AUC = 0.83; Acc = 0.747; Sens = 0.881; Spec = 0.617	No	NF
			Misawa et al. [52]	Proprietary	N/A	C3dNet	AUC = 0.87; Spec = 0.63 @ Sens = 0.90	No	NF
			Murthy et al. [53]	ISBI2014 challenge	Creation	CDDN	Acc = 0.8743	Yes	NF
			Pogorelov et al. [54]	CVC-ColonDB, CVC-ClinicDB, CVC-VideoClinicDB, Kvasir, Nerthus	Creation	Xception, VGG19, ResNet50, GAN	Acc = 0.909; Spec = 0.94	No	NF
		Bounding-box	Mo et al. [55]	CVC-ClinicDB, CVC-VideoClinicDB, CVC-ColonDB, CVC-EndoSceneStill	N/A	Faster R-CNN (VGG16)	Acc = 0.985; Prec = 1; Rec = 0.985; F1-score = 0.971; F2-score = 0.992	Comparison	NF
			Urban et al. [56]	Proprietary	On the flow	VGG16, VGG19, ResNet50	Acc = 0.964; ROC-AUC = 0.991	No	NF
		Semantic segmentation	Mohammed et al. [57]	ASU-Mayo	Creation, On the flow	Y-Net	Prec = 0.874; Rec = 0.844; F1-score = 0.859; F2-score = 0.850	Yes	NF
			Brandao et al. [58]	CVC-ClinicDB, ETIS-LARIB, ASU-Mayo	On the flow	FCN-AlexNet, FCN-GoogLeNet, FCN-VGG	Prec = 0.7361; Rec = 0.8631	Comparison	NF
Localization	Hybrid	Patch-based	Park et al. [59]	CVC-ClinicDB, ASU-Mayo	Patches	CNN	Prec = 0.6575; Rec = 0.8276	No	NF
		Bounding-box	Shin et al. [60]	CVC-ClinicDB, ETIS-LARIB, ASU-Mayo, CVC-VideoClinicDB	Creation	Faster R-CNN (Inception ResNet-v2)	Prec = 0.914; Rec = 0.803; F1-score = 0.833; F2-score = 0.815	Comparison	NF
			Yu et al. [61]	ASU-Mayo	Creation	3D-FCN	Prec = 0.881; Rec = 0.710	No	NF
			Zhang et al. [62]	ASU-Mayo, CVC-ClinicDB, ETIS-LARIB	On the flow	RYCO	Prec = 0.886; Rec = 0.716; Spec = 0.970	Yes	NF
		Bounding-box, semantic segmentation	Wang et al. [63]	CVC-ClinicDB, proprietary	N/A	SegNet	Sens = 0.9438; Spec = 0.9592	No	NF
		Feature extractor, patch-based	Billah et al. [64]	Mesejo, ASU-Mayo, CVC-ClinicDB, ETIS-LARIB	N/A	CNN	Acc = 0.9865; Sens = 0.9879; Spec = 0.9852	Comparison	NF
	End-to-end	Bounding-box	Mo et al. [55]	CVC-ClinicDB, CVC-VideoClinicDB, CVC-ColonDB, CVC-EndoSceneStill	N/A	Faster R-CNN (VGG16)	Prec = 0.862; Rec = 0.981; F1-score = 0.917; F2-score = 0.956	Comparison	NF
			Zheng et al. [65]	CVC-ColonDB, CVC-ClinicDB, ETIS-LARIB, proprietary	Creation	YOLO	Prec = 0.774; Sens = 0.740; F1-Score = 0.757; F2-score:0.747	No	NF
			Pogorelov et al. [66]	ASU-Mayo, proprietary	N/A	TensorBox, Darknet-YOLO	Acc TensorBox = 0.316; Acc Darknet-YOLO = 0.422	No	NF
		Classification	Urban et al. [56]	Proprietary	On the flow	VGG16, VGG19, ResNet50	Dice = 0.827 ± 0.003	No	NF
		Patch-based, classification, semantic segmentation	Pogorelov et al. [54]	CVC-ColonDB, CVC-ClinicDB, CVC-VideoClinicDB, Kvasir, Nerthus	Creation	Xception, VGG19, ResNet50, GAN	Acc = 0.946; Spec = 0.984	No	NF

(continued on next page)

Table 3 (*continued*)

Objective	Type	Approach	Authors	Datasets	Data augmentation	Network	Main results	Baseline	Code
Segmentation	Hybrid	Patch-based	Zhang et al. [67]	CVC-ColonDB	N/A	FCN-8s	Acc = 0.9754; Rec = 0.7566; Spec = 0.9881; Dice = 0.7014	Comparison	NF
	End-to-end	Semantic segmentation	Nguyen and Lee [68]	CVC-ClinicDB, ETIS-LARIB	Pixel level	Encoder-decoder	Dice = 0.889; IoU = 0.8935; Acc = 0.984	Comparison	NF
			Wichakam et al. [69]	CVC-EndoSceneStill	On the flow	Compressed FCN-8s	Prec = 0.8848; Rec = 0.7814; IoU = 0.6936; Dice = 0.9594	Yes	NF
			Wickstrøm et al. [70]	CVC-EndoSceneStill	On the flow	Enhanced FCN-8, enhanced SegNet	IoU = 0.767; Acc = 0.949	Comparison	Yes[a]
			Xiao et al. [71]	CVC-ClinicDB	N/A	DeepLab-v3+LSTM	IoU = 0.9321	Comparison	NF
			Zhou et al. [72]	ASU-Mayo	N/A	U-Net++	IoU = 0.3345	Yes	Yes[b]
			Bardhi et al. [73]	CVC-ColonDB, CVC-ClinicDB, ETIS-LARIB	On the flow	SegNet	Acc = 0.967	No	NF
			Brandao et al. [58]	CVC-ClinicDB, ETIS-LARIB, ASU-Mayo	On the flow	FCN-AlexNet, FCN-GoogLeNet, FCN-VGG	Prec = 0.7023; Rec = 0.5420	Comparison	NF
			Li et al. [74]	CVC-ClinicDB	Creation	FCN	Acc = 0.9698; Rec = 0.7732; Spec = 0.9905	No	NF
			Vázquez et al. [36]	CVC-EndoSceneStill	On the flow	FCN-8	IoU Polyp = 0.5160; MGA = 0.9677	Yes	Yes[c]

N/A: Not applicable; CNN: convolutional neural network; FCN: fully convolutional network; Prec: precision; Rec: recall; Sens: sensitivity; IoU: Intersection over Union; MGA: mean global accuracy; MCA: mean classes accuracy; Spec: specificity; FPPF: false positive per frame; TPR: true positive rate; FPR: false positive rate; NF: not found.

[a] https://github.com/Wickstrom/Thesis.
[b] https://github.com/MrGiovanni/UNetPlusPlus.
[c] https://github.com/jbernoz/deeppolyp.

provides a classified set of videos depending on the Boston bowel preparation scale, therefore not providing any polyp information and being excluded from Table 4. Mesejo et al. [93] also provide a labelled dataset of 76 videos of different lesions (serrated adenomas, hyperplastic lesions and adenomas), which is used by Billah et al. [64]. Since optical biopsy for polyp classification is out of the scope of the current review, the dataset has not been included in Table 4.

Regarding clinical variability, few datasets indicate the type of polyps included. The Paris classification [94,95] is a general framework for the endoscopic classification of superficial lesions of the oesophagus, stomach, and colon. Fig. 4 shows the different types of polyps, both in schematic view and actual endoscopic images. Pedunculated and sessile polyps are easier to detect than flat polyps and CAD systems to assist their detection would be more useful for gastroenterologists, but regretfully they are underrepresented in the public datasets [96].

4.2. Use of the datasets and data augmentation

Authors do not follow a standard methodology to distribute the dataset into training, validation and test sets, except for those using CVC-EndoSceneStill, because the distribution is provided by the dataset owner; or those following the rules of the MICCAI 2015 Automatic Polyp Detection in Colonoscopy Videos challenge. Works for detection use a greater number of images than those for segmentation, which might be because labelling frames is easier than manually segmenting polyps for ground truth creation (Fig. 5). Test sets are usually one or two degrees of magnitude smaller than the training sets. Due to this heterogeneity in training, validation and test sets, it is no easy to make a fair comparison of the methods and their reported metrics.

Data augmentation is the process whereby the training dataset is artificially increased in size, which in medical imaging is typically done with transformations that are applied to only the image in the case of detection (as each image only have a label that remains unaffected) or to the image and mask in the same way in the case of localization and segmentation. Augmentation methods commonly employ transformations such as rotations, reflections, and elastic deformations [97]. Data augmentation strategies are used by 24 out of the 35 analyzed papers (Table 3). The most common approach (8 works) is to enlarge the training set by creating new images through the application of transformations to the original images. Other authors (7 works) make data augmentation on the flow, i.e. transformations to the original image are randomly applied at training time, increasing the variability of the training set but without specifically creating new transformed images. Lastly, some other authors (8 works) train the models using patches extracted from the original images rather than using the full image. More recently, Nguyen and Lee [68] proposed a data augmentation approach at a pixel level for polyp segmentation.

To create new images, there is a wide variability on the transformations applied and their ranges (Table 5), as data augmentation is typically performed by trial and error and transformations are selected based on the imagination, time and experience of the researcher [98]. None of the authors justify the selection of neither the transformations nor the ranges. Few authors analyze the influence of data augmentation in the results. Vázquez et al. [36] compare results with and without different transformations, while Shin et al. [60] compare the influence of including more or fewer transformations. While the former identified that the combination of transformations leads to better results, the latter found that more augmentation does not guarantee better performance.

The different methods for data augmentation lead to a wide variability in terms of the actual training samples used (Fig. 5).

5. Comparison and discussion of methods

CNN [12,99,100] architectures are a type of neural networks which are specialized for data with grid-structured topology. CNNs are

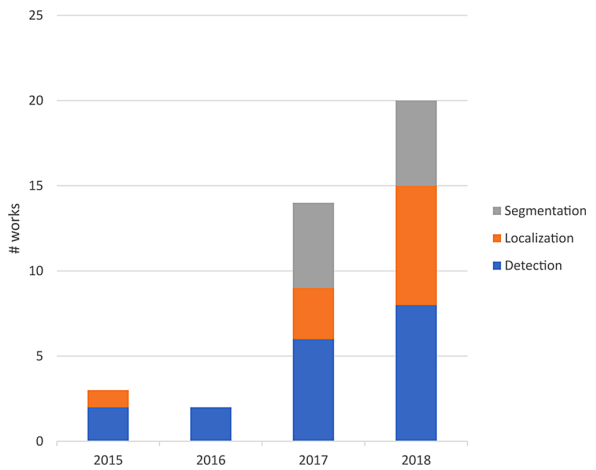


Fig. 2. Literature trends.

composed of different hierarchical stages that take advantage of local connections, shared weights, pooling layers and the use of many layers. The first set of stages may include a certain number of convolutional layers (namely two or three) followed by subsequent pooling layers. Convolutional layers exploit the local connections and shared weights by using the convolution operation instead of matrix multiplication and pooling layers subsample the data and merge similar features. Blocks of convolutional and pooling layers are then stacked to create a feature vector that represents the input data. Fully connected layers are latterly connected to this vector for the final object classification. One of the simpler and classical implementations of this stackable approach is VGG [76], where different stackable layers of 3×3 convolutions and maximum pooling layers are concatenated. However, this network presents a large number of parameters and slow convergence. He et al. [78] proposed the so called residual neural networks, which include skip connections so learnt filters are applied not to the final transformed image but to the residual over the input image instead, allowing deep neural networks to go deeper by mitigating the vanishing gradient problem and providing the first layers with larger scale gradients during backpropagation. Any CNN architecture can be extended into a FCN [101] by using a classification network as an encoder that is convolved over larger images producing spatially dense prediction tasks. However, these FCNs lacked the capability of generating fine shape delineation as high resolution reconstruction was calculated by interpolation. These capabilities were introduced by the addition of deconvolution layers, skip connections [80] and pooling indices [79]. This segmentation architecture was improved by using the fully convolutional DenseNet for image segmentation [102].

Based on the aforementioned works, different efforts have been followed during the last years for polyp detection, localization and segmentation, where deep learning-based approaches have proven to excel hand-crafted methods. An initial division of methods has been established by Bernal et al. [24]. Two different types of methods are of interest in this review: (1) hybrid methods that combine deep learning approaches with other hand-crafted methods and (2) end-to-end methods that use one single deep learning approach to obtain the result. A third type also mentioned by Bernal et al. [24], hand-crafted methods, lie out of the scope of this review. As secondary classification, we have grouped the approaches into five types, depending on their use of the deep learning network (Fig. 6):

1. Feature extractor. Deep learning architectures are used for automatically creating a feature vector, instead of the manual extraction of features. The computed vector is afterwards the input to a classical classifier, such as support vector machine (SVM) or distance-weighted discrimination (DWD) classifier [103], usually more robust than SVM. These methods are therefore always hybrid, as deep learning is combined with classical classifiers.
2. Classification. A classification network is used to label an image as containing a polyp or not, without position information of the polyp.
3. Patch-based. The method uses image patches or tiles and the presence of polyp is obtained for each patch. Location of the polyp might be obtained based on the patch location.
4. Bounding-box. The method provides the location of the polyp through a bounding-box (coordinates of the upper right corner, height and width), generally using a regression layer.
5. Semantic segmentation. Each pixel of the image is labelled as polyp or background. Networks based on encoder-decoder blocks are usually selected. The first half of the layers encode the image description highlighting the discriminative features for the entrusted task, while the second half is responsible for mapping the low-resolution encoding into full input resolution feature maps. These FCNs can be initialized from the weights of a classification network (encoder) that acts as a feature extractor.

All papers have been categorized using these primary and secondary classifications. While the primary classification is mutually exclusive, some works falls in more than one group of the secondary classification. Fig. 7 shows the distribution of papers. Since tasks and primary

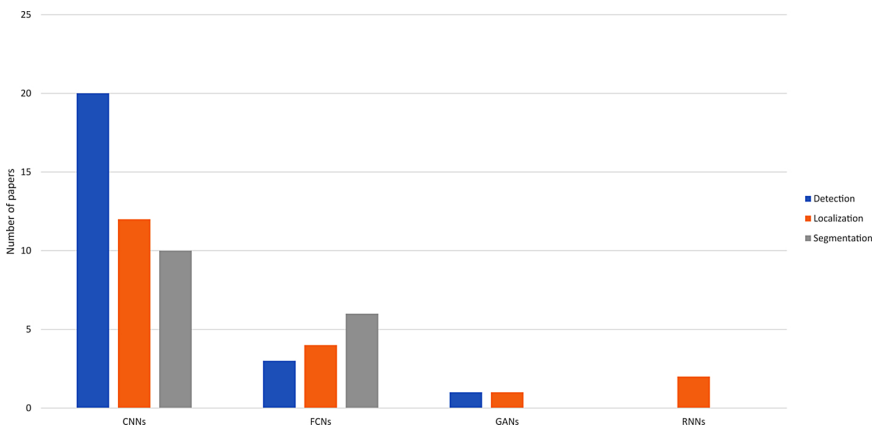


Fig. 3. Networks for each of the tasks. CNNs: Convolutional neural networks; FCNs: fully convolutional networks; GANs: generative adversarial networks; RNNs: recurrent neural networks.

Table 4
White-light datasets for polyp detection, localization and segmentation. Main natural image datasets are also included for comparison.

Dataset	Content	Ground truth	Delineation	Resolution	Type of polyps[a]	# patients	# items	Defined datasets	Training set	Validation set	Test set	Works[b]
CVC-EndoSceneStill [36]	WP images	BM (border, polyp, lumen and specular lights classes)	Manual by experts	500 × 574; 384 × 288	N/A	36	912	Yes	547	183	182	5
CVC-ColonDB [22]	WP images	BM (polyp)	Manual by experts	500 × 574	N/A	15	300	No	–	–	–	12
CVC-ClinicDB [21]	WP images	BM (polyp)	Manual by experts	384 × 288	0-Ip, 0-Is, 0-IIa, 0-IIb	23	612	No	–	–	–	20
ETIS-LARIB [24]	WP images	BM (polyp)	Manual by experts	1255 × 966	N/A	44 seqs	196	No	–	–	–	10
ASU-Mayo Clinic [88]	Video (WP and WO frames)	BM (polyp) (only training set)	Manual by experts	712 × 480; 856 × 480; 1920 × 1080	N/A	N/A	38 videos (8,591 WP; 27,979 WO)	Yes	20 videos (4,278 WP; 14,718 WO)	–	18 videos (4,313 WP; 13,261 WO)	22
CVC-VideoClinicDB [89,87]	Video (WP and WO frames)	BM (polyp) (only training set)	Elliptical approximation by experts	384 × 288	0-Is, 0-Ip, 0-IIa	18 seqs	36 videos	Yes	18 videos (11,954 frames)	–	18 videos (18,733 frames)	6
Kvasir [90]	Multi-Class Image Dataset, containing the polyp class	Label	N/A	from 720 × 576 up to 1920 × 1072	N/A	N/A	500 images in polyp class	No	–	–	–	3
ImageNet [81,82]	Natural images	Label and BB	Manual by annotators in Amazon Mechanical Turk	Variable	N/A	N/A	14,197,122	Yes	Up to 1.2 million	≈ 50,000	Up to 150,000	12
MSCoco [83,84]	Natural images	BM (171 classes), BB, person keypoints, captions	Manual by annotators in Amazon Mechanical Turk	Variable	N/A	N/A	330,000	Yes	Up to 118,287	Up to 5,000	Up to 40,670	1
Pascal VOC [85,86]	Natural images	BM (20 classes), BB	Manual by annotators in Amazon Mechanical Turk	Variable	N/A	N/A	19,041 images 40,657 instances[c]	Yes	9,477 images 20,250 instances	9,564 images 20,407 instances	N/A	2

WP: With polyp; WO: without polyp; BM: binary mask; BB: bounding-box; seqs: sequences; N/A: not available.
[a] According to Paris classification.
[b] Total add more than retrieved works because several authors use more than one database.
[c] Only training and validation sets.

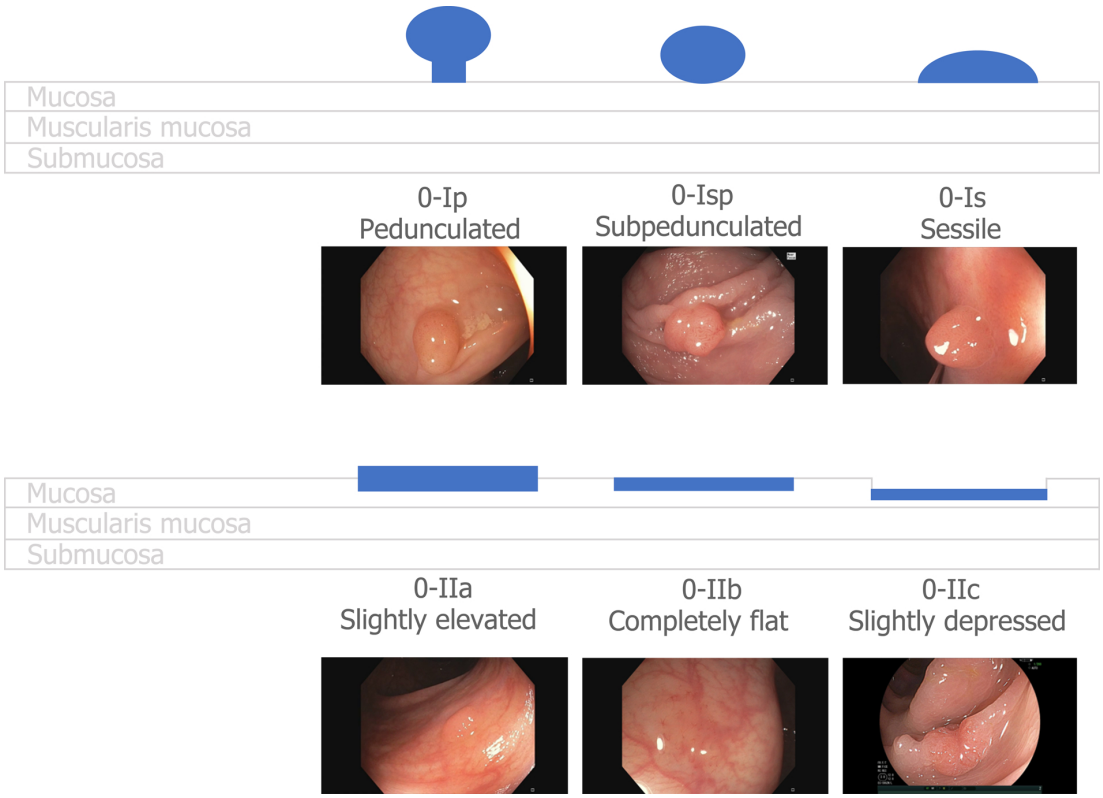


Fig. 4. Paris classification. Adapted from [24,95].

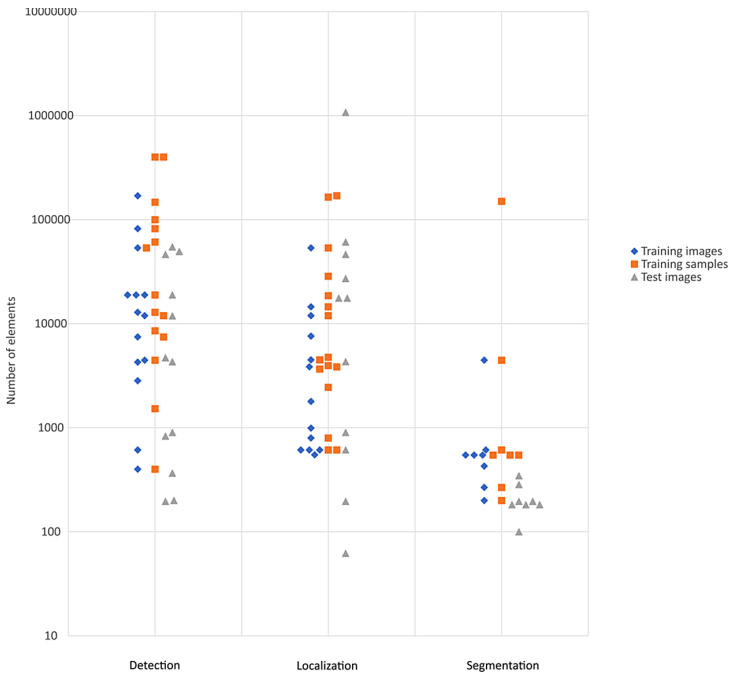


Fig. 5. Number of elements used for training and test, categorized by the main task of the work. 'Elements' refers to images for training and test sets as well as training samples, which might differ from training images if patches are extracted or data augmentation is applied.

Table 5
Transformations for data augmentation on full images.

Task	Work	Creation/flow	Rotation	Translation	Sampling	Zoom	Shearing	Warping	Flip	Gaussian noise	Contrast	Brightness	Crop	Increment
Detection	Aksenov et al. [50]	Creation	–	–	–	–	–	–	–	Yes	–	Yes	–	×1[a]
	Mohammed et al. [57]	Creation, on the flow	(10°,350°)	(−10, 10)	–	(1, 1.3)	–	–	H/V	–	–	–	Yes	×2
	Murthy et al. [53]	Creation	90°, 180°, 270°	–	–	–	–	–	H	–	–	–	–	×9
Localization	Shin et al. [60]	Creation	90°, 180°, 270°	–	–	–	–	–	H/V	–	–	–	–	×6
			90°, 180°, 270°	–	–	10%	10%, 30%,50%	–	H/V	–	–	–	–	×30
			90°, 180°, 270°	–	different resolutions	10%	10%, 30%,50%	–	H/V	–	–	Yes	–	×48
	Zhang et al. [62]	On the flow	Yes	–	–	–	–	–	H/V	Yes	Yes	Yes	–	×1
	Zheng et al. [65]	Creation	90°, 180°, 270°	–	–	–	–	–	–	–	–	–	–	×3
	Yu et al. [61]	Creation	90°, 180°, 270°	(−3, 3)	(−10, 10)	–	–	–	–	–	–	–	–	×22
Segmentation	Wichakam et al. [69]	On the flow	Up to 180°	(0, 20%) H/V	–	(-0.8, 1.2)	(0, 0.2)	–	H/V	–	–	–	–	×1
	Wickstrøm et al. [70]	On the flow	(-90°, 90°)	–	–	(0.8, 1.2)	(0, 0.4)	–	–	–	–	–	Yes	×1
	Bardhi et al. [73]	On the flow	Yes	Yes	0.5 dropout	–	Yes	–	H/V	Yes	–	–	Yes	×1
	Li et al. [74]	Creation	Yes	H/V	–	–	–	–	–	Yes	Yes	–	–	×333
	Vázquez et al. [36]	On the flow	(0°, 180°)	–	–	(0.9, 1.1)	(0, 0.4)	(0, 10)	–	–	–	–	–	×1
Detection and localization	Pogorelov et al. [54]	Creation	20° steps	–	–	–	–	–	H	–	–	–	–	×35
	Urban et al. [56]	On the flow	(0°, 90°)	–	–	–	Yes	–	H/V	–	–	–	–	×1
Detection and segmentation	Brandao et al. [58]	–	–	–	–	–	–	–	Random	–	–	–	–	×1

[a] They modify 40% of the images to increase variability, but do not create more new training examples. H: horizontal; V: vertical.

classification are mutually exclusive, six Venn's diagrams are necessary to show the overlap between the categories of the secondary classification. Venn's diagrams have been created with InteractiVenn [104].

End-to-end and hybrid methods have similar proportion in detection and localization (close to 50% each), but in segmentation, the end-to-end methods vastly surpass hybrid ones (9 vs 1). Regarding the secondary categorization, classification and patch-based ranked equally for the detection task; while bounding-box and semantic segmentation are the preferred approaches for localization and segmentation, respectively. Hybrid, patch-based methods and end-to-end, semantic segmentation methods are the most usual combinations.

Fig. 8 shows how the different types of methods have had presence along the years analyzed in this systematic review. In summary, the tendency goes towards the use of end-to-end methods over hybrid ones, as deep learning is gaining more and more capabilities to address complex problems as a whole, rather than being used as a component to codify the image into features that are afterwards further analyzed. Similarly, semantic segmentation has raised interest over the last two years as it comprises a straightforward and seamless method for polyp identification.

The use of data augmentation and pre-trained networks in the retrieved works have also been analyzed. Fine-tuning is a well-known alternative to training a network from scratch when the labelled training data is limited. In this case, pre-trained networks on a large labelled dataset from a different application is used as starting point [45]. Fig. 9 shows the number of training samples after applying data augmentation, if any, of each method according to the type of approach. There is no clear trend in the use of data augmentation strategies combined with pre-trained models in the considered approaches. It might be expected that using pre-trained models would be linked to a lower number of training samples, but findings do not show so.

Lastly, it is worth mentioning that all retrieved works apply supervised learning, relying its training on a labelled set of images. Other learning approaches already applied in different medical fields, such as unsupervised learning [105] or few-shot learning [106,107] have not been applied for polyp detection, localization or segmentation yet. Unsupervised learning might be more difficult to apply because the polyp area is usually a small portion of the image, while the rest presents a high level of similarity, as there is also healthy mucosa in an image labelled as with polyp.

In the following sections, methods are briefly described, grouped accordingly to the primary and secondary classifications. When methods fall into two groups of the secondary classification, they are each indicated in a different paragraph.

5.1. Methods for polyp detection

5.1.1. Hybrid methods

5.1.1.1. Feature extractor and patch-based. Several authors have compared the use of hand-crafted advanced features with simple fine-tuned classification CNNs, and in all cases the CNN-based approaches overcame the manual feature extraction. Shin et al. [42] demonstrate that features obtained with a classification CNN perform much better than hand-crafted features even when using state-of-art histogram of oriented gradients descriptors. They employ a basic architecture of three convolution:max pooling layers followed by a fully connected layer with 256 neurons image patches to feed an SVM.

Similarly, Taha et al. [41] perform a comparative testing of a shallow classifier fed either with hand-crafted methods or with features extracted using a pre-trained classification CNN. These authors employ AlexNet to obtain a 1,000-dimensional feature vector that is the input to an SVM.

In these works, images patches are used, so these methods can also be classified as patch-based detection but without localization information, as location of patches is not considered.

5.1.1.2. Patch-based. In this case, Yuan et al. [44] propose a 2-stage method. In first place, candidates are detected by the analysis of edges. Patches are then cropped around the polyp candidates and the resulting candidate patches are analyzed by a classification network based on AlexNet.

Tajbakhsh et al. in various works [43,45–47] extend this method by using a pre-detection stage based on edge maps and voting schemas to extract all suitable candidates that are classified by AlexNet, either trained from scratch or fine-tuned, depending on the work. Oriented patches are extracted based on shape features, conforming the set of candidates.

5.1.1.3. Classification and patch-based. In this case, Axyonov et al. [48] combine image contrasting and the K-means-with-connectivity-constraint segmentation method to identify regions with similar pixels, which are then classified into polyp or non-polyp region using AlexNet as classifier.

5.1.2. End-to-end methods

5.1.2.1. Classification. Akbari et al. [49] use a CNN made of four convolutional and pooling layers plus two fully connected layers. This CNN is combined with binarized weights and kernels to reduce the CNN size, so it is suitable for implementation in portable medical hardware with limited memory.

On the other hand, Aksenov et al. [50] combine various classification networks through classifier assembling to get higher accuracies, basing the final result on the average result of the three ensembled models. Each model presents a different number of layers as well as different configurations for the filters.

Itoh et al. [51] aim at detecting polyps as a prior stage to polyp size estimation. In their approach, they use a 3D CNN (C3dNet or C3D) [108] that exploits both the spatial structure and the temporal features present in colonoscopy videos by using 3D convolutional filters and 3D pooling layers. The input is a sequence of 16 consecutive frames for which the CAD system provides an output probability of being a sequence with or without polyp. The same network is used by Misawa et al. [52], who focused their effort on a clinical trial to measure the sensibility of the C3dNet.

A different approach is followed by Mo et al. [55], who employ a Faster R-CNN with VGG16 as backbone. In this case, they use an approximately joint optimization, which takes a mini-batch as input and optimizes both the classification and regression losses at the same time. The classification tail allows for detecting polyps in a frame by giving a probability level.

Classification networks have also been proven successful even on colon cancer screening programmes [56]. In this case, they test pre-trained VGG16, VGG19 and ResNet50 with a final binary classification layer, as well as custom, trained-from-scratch CNNs for polyp detection without further processing. Images were classified on polyp or non-polyp classes.

Lastly, Murthy et al. [53] introduce a cascaded deep decision network (CDDN) for classification. In the first learning stage, samples are classified by a pre-trained network. In the second training stage, samples classified with high confidence are discarded to place efforts on the most challenging samples. This process is repeated on successive stages, using previous stage's features, and obtaining a better separation over samples. In this particular case, a 2-stage network is proposed.

5.1.2.2. Classification and semantic segmentation. Pogorelov et al. [54] compare three detection approaches: (1) hand-crafted global features and a logistic model tree classifier; (2) fine-tuned well-known architectures, such as Xception [109], VGG19 and ResNet50; and (3) a pixel-wise segmentation GAN with a threshold on the number of positively labelled pixels. They use different training and testing sets for the three suggested methods and find that the GAN approach performs better.

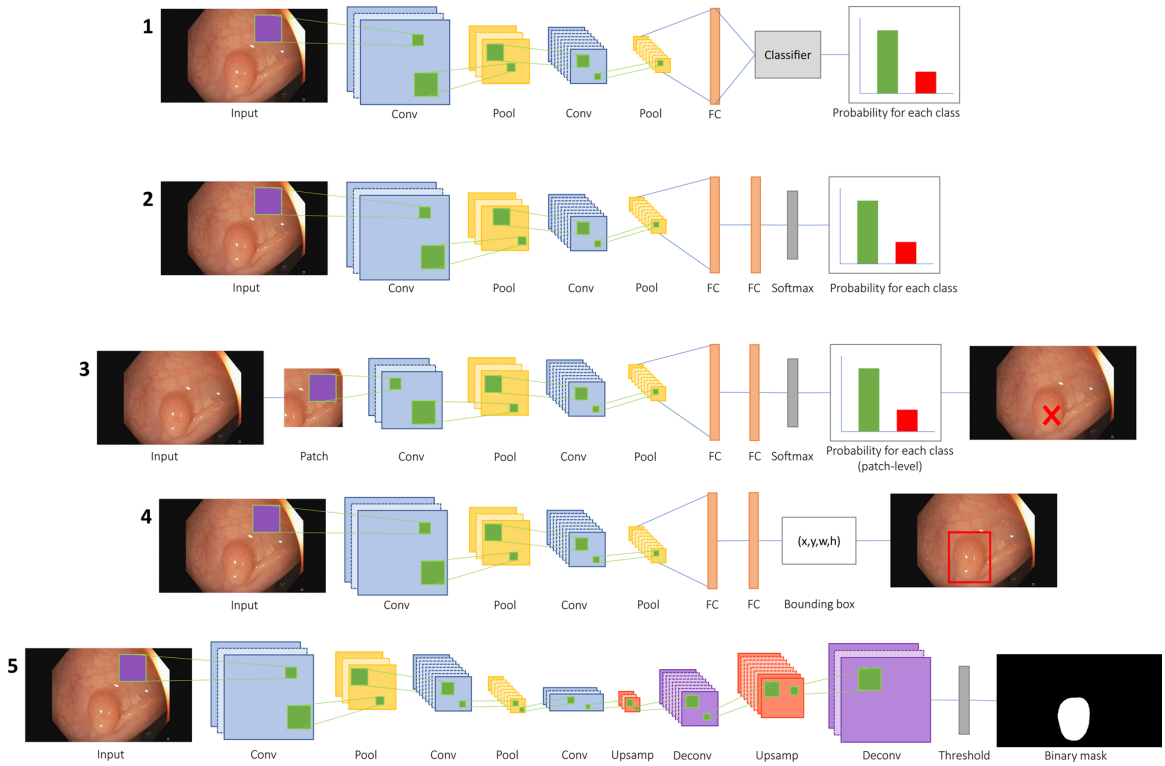


Fig. 6. Schematic representation of the five considered approaches in this review. From top to bottom, (1) feature extractor, (2) classification, (3) patch-based, (4) bounding-box, and (5) semantic segmentation. Each type of layer is represented by a different colour: convolutional layer (conv); pooling layer (pool), fully connected layer (FC), upsampling layer (upsamp) and deconvolutional layer (deconv). The receptive field is marked with a green square. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

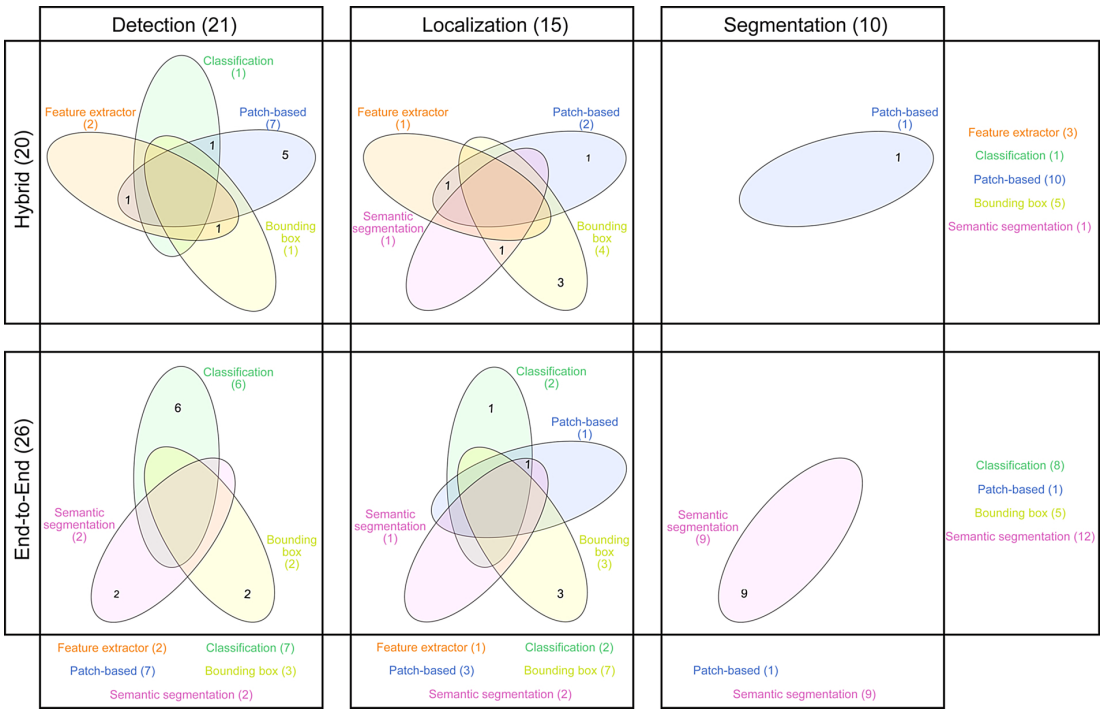


Fig. 7. Categorization of works per tasks.

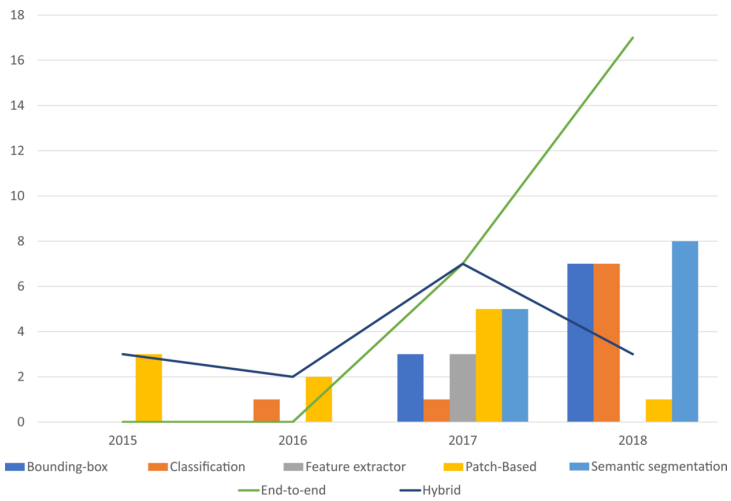


Fig. 8. Trends of the different approaches along the years.

5.1.2.3. Semantic segmentation. Mohammed et al. [57] design the Y-Net, which combines two encoders (both use VGG19 as backbone; pre-trained for encoder 1, while trained-from-scratch for encoder 2) and one decoder to produce a pixel-wise segmentation. Detection is considered when the Intersection over Union of the predicted result and the ground truth is greater than 0.90. In contrast, although Brandao et al. [58] also base detection on the semantic segmentation performed with different FCNs, they consider any degree of overlap between the predicted result and the ground truth. Therefore, not setting an overlap minimum might lead to better metrics in comparison to Mohammed et al. [57], who set their overlap threshold at 0.9.

5.1.3. Comparison of detection methods

Although it is difficult to compare results because of the different datasets, we focus firstly on accuracy as an important indicator to compare between models and approaches (Table 6). With this in mind, Mo et al. [55] deliver the best results in detection. These results are obtained using three different test sets (CVC-ClinicDB, CVC-ColonDB and CVC-EndoSceneStill) and averaging the results obtained for each one. Urban et al. [56] also achieve a high accuracy. In this case, the accuracy is identical in both cross-validation and independent test set of a proprietary dataset. Remarkable results in a test set from ASU-Mayo Clinic database are also those presented by Yuan et al. [44], who achieve an accuracy of 0.9147, with a recall of 0.9176. Akbari et al. [49] also use the same ASU-Mayo Clinic database but with a distinct test set. In this work, the accuracy obtained is 0.908, and the recall, precision and specificity are 0.6838, 0.7434 and 0.9497, respectively. Confidence intervals (CI) are indicated when provided by authors. This applies also to Sections 5.2.3 and 5.3.3.

5.2. Methods for polyp localization

5.2.1. Hybrid methods

5.2.1.1. Patch-based. Park et al. [59] use an architecture composed of three convolutional layers and three max-pooling layers to obtain a 60-dimension feature vector for each patch obtained at three different image scales. The resulting 180-dimension feature vector is used to classify the centre pixel of the patch either as polyp or non-polyp, through a fully connected network with 256 hidden nodes. A probability map is created based on these classified pixels. This map is smoothed with a 5×5 Gaussian filter and 9×9 median filter. Afterwards, it is thresholded, setting to 0 those pixels with probability lower than 0.65. Lastly, connected components of non-zero regions are identified, and the polyp centre is obtained by calculating the centre of mass of each connected component.

5.2.1.2. Feature extractor and patch-based. Billah et al. [64] use a 10-layer CNN to extract features from the last fully connected layer, which are then combined with wavelet features and all of them fed into an SVM for classification into polyp or non-polyp. Patches corresponding to a sliding window are the input, so location of the polyp is found by averaging regions with higher probabilities of being polyp.

5.2.1.3. Bounding-box. The method of Shin et al. [60] combines a region proposal network (RPN), a detector and post-learning approach in the so-called Faster R-CNN, using Inception ResNet-v2 [110] as backbone. Out of an image, the RPN proposes rectangular candidate regions that are the input to the detector, which classifies them into containing or not containing a polyp. The detection is further improved by the post-learning approach, based on false positive learning and off-line learning.

Yu et al. [61] integrate temporal information into the model. Their 3D-FCN is capable of learning more representative spatio-temporal features from colonoscopy videos and hence has more powerful discrimination capability. This 3D network consists of a 3D extension of a 2D fully convolutional segmentation network that uses a 16 frames video entrance to extract the temporal features. An offline 3D-FCN is firstly trained, which is combined with an online 3D-FCN incrementally updated for each input video to remove false positives. Outputs of these two 3D-FCNs are combined to obtain the final detection results.

Moreover, Zhang et al. [62] propose an evolution of a YOLO detection network [111], ResYOLO. The previous detection output of the network is integrated into the following prediction to assure prediction regularization. Besides, temporal information is incorporated as an online object tracker. They also analyzed that the inclusion of temporal information on the network improves the detection rates.

5.2.1.4. Bounding-box and semantic segmentation. SegNet is well-known for semantic segmentation of natural images, so it is the network selected by Wang et al. [63]. Although SegNet provides a pixel-wise labelling in the form of a probability map, there is a post-processing stage that transforms it into the corresponding bounding-box for the polyp class.

5.2.2. End-to-end methods

5.2.2.1. Bounding-box. As mentioned before, Mo et al. [55] use a Faster

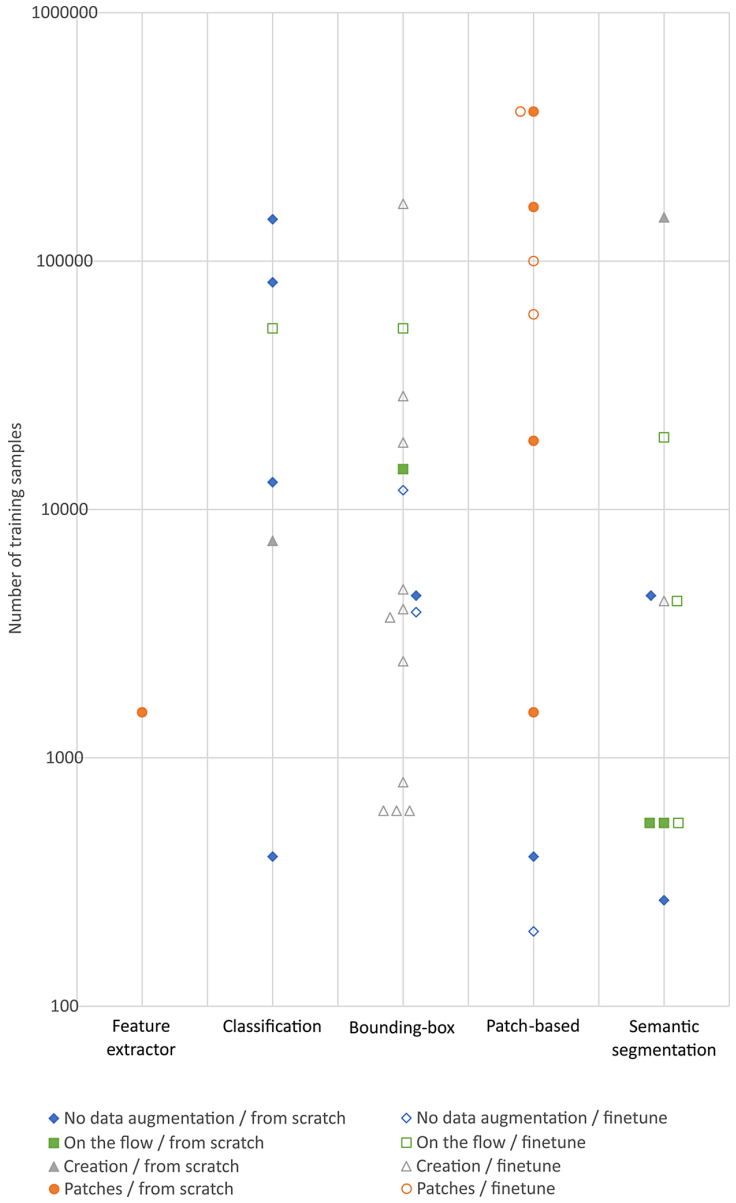


Fig. 9. Training samples per approach and type of data augmentation. Solid items correspond to methods where networks are trained from scratch, while hollow items correspond to fine-tuned networks. In the horizontal axis, the different approaches are indicated. The series corresponds to the different strategies for data augmentation.

Table 6
Summary of detection methods results.

Work	Accuracy	Precision	Recall	AUC	Specificity
Mo et al. [55]	0.985	1.000	0.985		
Urban et al. [56]	0.964			0.974	
Yuan et al. [44]	0.915		0.918		
Akbari et al. [49]	0.908	0.743	0.684		0.950

AUC:area under the curve

R-CNN with VGG16 as backbone. In this case, the regression tail provides the coordinates of the bounding-box indicating the location of the polyp.

Urban et al. [56] compare models trained from scratch to pre-trained models (VGG16 and VGG19 with a final regression layer), obtaining a higher Dice for the latter when ground truth and predicted bounding-boxes are compared.

On the other hand, YOLO is selected by Zheng et al. [65] as the detector network, without further modifications on the network and proceeding only with a fine-tuning of a pre-trained model.

Pogorelov et al. [66] compare two different approaches, TensorBox and Darknet-YOLO, both aiming at detecting objects in images. On one hand, TensorBox avoids multiple detections of the same object by using an RNN with an LSTM. On the other hand, Darknet-YOLO is based on a CNN, therefore encoding contextual information about classes as well as their appearance. This results in a better generalization of objects' representation. In both cases, the methods return sets of rectangles

marking possible polyp locations together with corresponding location confidence values.

5.2.2.2. Classification and patch-based. Pogorelov et al. [54] use sliding windows to feed the classification models (hand-crafted global features, fine-tuned networks and GAN approach grounded on V-GAN [112] modified by adding an activation layer to generate a per-pixel image segmentation, so detection is based on a minimum number of activated pixels) and then reconstruct a coarse localization map by grouping-back the processed patches.

5.2.3. Comparison of localization methods

Out of all localization methods (Table 7), Billah et al. [64] report the highest metrics. In this case, colour wavelet features and CNN features are combined, so it is not possible to determine which type of feature has a greater influence on the result. Besides, they do not report the testing data set, neither in terms of number of images nor their origin, which makes it more difficult to compare results. On the contrary, Mo et al. [55] clearly indicate the testing datasets (CVC-ClinicDB, CVC-ColonDB and CVC-EndoSceneStill), reporting a mean recall which is comparable to the previous work, but providing a more solid evidence and reproducibility of results. Nevertheless, no further works use the same validation set. In this regard, Yu et al. [61] and Zhang et al. [62] do follow the rules of the MICCAI 2015 Automatic Polyp Detection in Colonoscopy Videos challenge in terms of testing dataset. This way, they provide a fair and straightforward comparison to other methods participating in the challenge [24]. Although none of them outperform ASU, the winning method, in precision, they do provide better recall, F1-score and F2-score, showing a more balanced performance.

Table 7
Summary of localization methods results.

Work	Accuracy	Recall	F1-score	F2-score
Billah et al. [64]	0.987	0.988		
Mo et al. [55]		0.981		
Yu et al. [61]	0.881	0.710	0.786	0.739
Zhang et al. [62][a]	0.886	0.716 (0.703, 0.730)	0.792 (0.780, 0.804)	0.744 (0.732, 0.757)
ASU [24]	0.935	0.611	0.739	0.657

[a] 95% confidence intervals are provided between brackets.

5.3. Methods for polyp segmentation

5.3.1. Hybrid methods

5.3.1.1. Patch-based. The focus of Zhang et al. [67] is placed on the use of texton-based spatial features for detailed classification that is used to remove false positives based on local textural analysis. In this case, results of the FCN-8s [101] are used to extract the image region proposals. These regions are refined using texton-based patch representation, which is followed by a random forest classifier to provide the final segmentation.

5.3.2. End-to-end methods

5.3.2.1. Semantic segmentation. Nguyen and Lee [68] take an encoder-decoder model as basis and then produce the polyp segmentation using a model combination by training the encoder-decoder model with three different resolutions databases.

Besides, Wichakam et al. [69] present a compressed FCN that reduce the number of parameters of the feature vector extracted by the network to minimize the computational time showing faster convergence and increased performance for polyp segmentation. The model is compressed by substituting two $7 \times 7 \times 4096$ convolutionalized layers by one $7 \times 7 \times 512$ convolutionalized layer, reducing the number of trainable weights in a significant manner.

The proposal of Wickstrøm et al. [70] is to enhance two traditional encoder-decoder networks (FCN-8s and SegNet, using both VGG16 as encoder) by including batch normalization after each layer and dropout after the three central encoders and decoders. They also analyze uncertainty and interpretability of the models.

On the other hand, Xiao et al. [71] combine LSTM with DeepLab-v3 in parallel. While the latter learns and extracts polyp features thanks its wide field-of-view and higher resolution, the former aims at preserving the information of the polyp location using the information stored in the memory cells These are regulated through the input, forget and output gates.

The U-Net has been modified into the U-Net++ by Zhou et al. [72]. The main difference is the inclusion of nested dense convolutional blocks that bridge the semantic gap between the feature maps of the encoder and decoder prior to fusion.

Bardhi et al. [73] directly use SegNet and train it from scratch on different datasets, while Brandao et al. [58] transform several traditional classification backbones for the segmentation networks into FCNs and prove that VGG16 backbone works better than GoogLeNet and AlexNet.

Similarly, Li et al. [74] propose an FCN and U-Net based segmentation network. The encoder is composed of 8 convolution layers, 8 rectified linear unit (ReLU) layers and 5 pooling layers, while the decoder includes 5 deconvolution layers, 5 concat layers, 6 convolution layers and 6 ReLU layers. Both stages are linked using skipping connections.

Finally, Vázquez et al. [36] provide an exhaustive benchmark showing that FCNs outperform previous results. They implement the FCN-8s architecture and test the influence of data augmentation and number of classes to be segmented on the network performance.

5.3.3. Comparison of segmentation methods

Table 8 summarizes the results of the segmentation methods. The highest Intersection over Union (IoU) values are obtained by Xiao et al. [71], in 345 images from CVC-ClinicDB, and Nguyen and Lee [68], reporting on the ETIS-LARIB dataset. In both cases, mean IoU is reported, so both polyp and background classes are considered, which explains to a great extent those values close to 1. This fact can be clearly seen in the work by Wickstrøm et al. [70]. IoU for the polyp class is 0.587 but mean IoU is equal to 0.767 thanks to the value of IoU for the background, as high as 0.946.

The comparison of methods that use CVC-EndoSceneStill is fair and straight-forward, as the dataset owners provide its division into training, validation and testing sets. Therefore, Vázquez et al. [36], Wichakam et al. [69] and Wickstrøm et al. [70] report results on the same 182 images. Regretfully, and although the CVC-EndoSceneStill benchmark provides a set of metrics (IoU and accuracy), not all authors calculate them. Wichakam et al. [69] follow instead the metrics given by the Pascal VOC challenge [85]. Vázquez et al. [36] report slightly higher values than Wickstrøm et al. [70] in terms of accuracy but it is on the contrary when IoU is considered. In this regard, Wichakam et al. [69] report an intermediate value of IoU.

Bardhi et al. [73] do not clearly state the division into training and testing datasets; therefore results, although showing high values, should be interpreted carefully. The only hybrid method for segmentation [67] obtains comparable results to the rest of end-to-end methods.

In the segmentation task, it is important to remark the influence of including the background class when calculating metrics. Since background usually means the largest area within a frame in comparison to the polyp class, background affects the results by increasing the metric value even when the segmented result is poor. This issue is further discussed in Section 6.

5.4. Advantages and disadvantages

Table 9 gathers the main advantages and disadvantages of the categories on which the works have been classified.

5.5. Loss functions

The final goal against which the network is optimized is completely defined by the loss function. While traditional loss functions for classification (such as negative log-likelihood – NLL – or cross entropy), regression (i.e. L1-loss or mean squared error – MSE) and distribution matching tasks (such as the Kullback-Leibler – KL – divergence) might be generalizable enough for most of the problems, the selection of the loss function is more relevant for complex problems such as the segmentation of unbalanced classes [113]. In these cases, an inappropriate formulation of the loss function might cause that the network converges into a minima where the task is not achieved because the network does not behave as expected. In this review, only 22.86% of the analyzed works present an analysis on the selected loss function and less than 15% of the authors use a custom loss function. These works are briefly commented below.

The loss in the Faster R-CNN used by Mo et al. [55] consist of a classification loss and a bounding-box regression loss, using parameterized coordinates to minimize the influence of scales during training. Besides, Zhang et al. [62] employ a loss function comprised of loss for grids labelled as object (polyp) and non-object. Furthermore, other three works include the Dice coefficient. On one hand, Wichakam et al. [69] optimize the network with a custom loss function simply computed as $1 - Dice$. On the other hand, while Zhou et al. [72] combine this coefficient with binary cross-entropy, Mohammed et al. [57] use the weight binary entropy instead.

When deciding the loss function for detection, it is important to know beforehand the proportion of polyp and non-polyp images in the training set. If it is balanced, traditional loss functions might be enough. Otherwise, it would be advisable to use a loss function intended to overcome the unbalance towards the positive class, such as the weighted binary cross entropy. We also agree on including overlap measures such as the Dice coefficient in the loss function for segmentation tasks when classes are highly unbalanced, as this type of functions have been proved to be more robust [113]. Since segmentation can be done as a pixel-wise binary classification of unbalanced classes, comments for detection can also be applied here.

6. Metrics

There is a wide variety of metrics found in this systematic review. In all cases, metrics are intended to compare the prediction of the method, either for detection, localization or segmentation, against the ground truth, which might be a label, a bounding-box or a binary mask. In this regard, many metrics are calculated based on the confusion matrix and its four basic elements (Fig. 10):

- True positives (TP): number of polyp items correctly predicted as polyp.
- True negatives (TN): number of non-polyp items correctly predicted as non-polyp.
- False positives (FP): number of non-polyp items incorrectly predicted as polyp.
- False negatives (FN): number of polyp items incorrectly predicted as non-polyp.

Table 8
Summary of segmentation methods results.

Work	Accuracy	IoU	F1-score	F2-score
Xiao et al. [71]		0.932		
Nguyen and Lee [68]		0.889		
Wickstrøm et al. [70]	0.949	0.767	0.786	0.739
Vázquez et al. [36]	0.968	0.516	0.792	0.744
Bardhi et al. [73]	0.967			
Wichakam et al. [69]		0.694	0.739	0.657

It is also important to point out that some authors compute the confusion matrix at a frame or image level, while others compute it at a pixel level. This mainly depends on the main objective of the work, being more usual to calculate the confusion matrix at frame level for detection and localization, and at pixel level for segmentation.

Table 10 compiles all metrics found during the analysis, providing the mathematical calculation and alternative names. The calculation of the Dice coefficient and IoU metrics based on the confusion matrix can be used when calculating the confusion matrix at pixel level by comparing two binary masks. These metrics are also more appropriate as they reduce the unbalance due to the large value of true negatives. Two definitions for accuracy have been found in this review. On one hand, only TP are considered in the numerator. This definition can be interpreted as class accuracy, as the sum of accuracies for all classes would reach 1 in an ideal situation. On the other hand, both TP and TN are considered in the numerator. In both cases, the denominator remains unchanged, being the total sum of elements (TP + TN + FP + FN). In the case where the formula is not explicitly provided, accuracy has been considered as correct classified items (TN + TP) divided by the total number of elements.

Table 11 summarizes the metrics employed for the different considered tasks. There is no standard to follow in terms of reporting metrics; therefore for one single task authors report different metrics. Overall, recall is the most used metric followed by precision, accuracy and F1-score in similar proportion. As for each of the analyzed tasks, recall at frame level is the predominant metric for detection and localization, while IoU is the most used metric for segmentation, closely followed by accuracy at pixel-level.

For detection and localization, recall stands out. This metrics penalizes a high number of FN but does not consider FP. In a clinical setting, both parameters are equally important, as if gastroenterologists are warned unnecessarily, the exploration time might be increased without benefits and they would eventually pay no attention to the CAD system ringing false alarms. Therefore, we consider a more suitable metrics the use of the F1-score, where importance of recall and precision is balanced.

Regarding segmentation metrics, it is important to point out that their selection must consider the properties of the segmentation under evaluation, so when the segment size is smaller than the background (less than 5% of the background in any of the axis, as it is usually the case in polyp segmentation), metrics based on the confusion matrix elements are not the more suitable and it is recommended to substitute them by distance metrics [114]. Therefore, when background is not considered in the calculation of the metric, a smaller increment in its value might be more significant than the same increment in a metric considering the background. Bearing this in mind, reporting accuracy, where true negatives corresponding to the background are included, does not seem to be the most adequate one, despite being one of the most common.

It is noteworthy to remark that traditional object detection metrics have not been found in any of the analyzed works. These metrics, such as average precision (AP) or mean average precision (mAP), measure the average precision of a model for a specific IoU and are commonly used on computer visions challenges [83,115]. This is relevant as these metrics provide a single value estimation on the performance of object detection models. However, although mAP can be used for computational performance estimation, it does not reflect the clinical performance due to its sensitivity for small objects and integration of the different. Thus, in clinical gastrointestinal works where mAP is used to measure the overall performance of the method, this has to be complemented with further analysis such as precision/recall curve analysis [116].

In general terms, metrics are reported for the overall testing set. Only few authors report metrics in a detailed manner, considering polyps characteristics. Misawa et al. [52] analyze the percentage of the video for each of the 50 polyps, for which their Paris classification, size,

Table 9
Advantages and disadvantages of each category of methods.

Category	Advantages	Disadvantages
End-to-end	1. Automatic learning of relevant features 2. Superior performance over hand-crafted methods 3. High degree of automation	1. Requires a large dataset for training 2. Selection of network, dataset and hyperparameters might highly influence the performance
Hybrid	1. Combines useful information of well-known hand-crafted features 2. More convenient with small datasets	1. Selection of features based on the researcher experience and knowledge
Feature extractor	1. Combine useful information of well-known hand-crafted features 2. Using pre-trained networks without fine-tuning does not require any labelled data	1. Not tune the network to the target dataset, potentially leading to suboptimal or even negative transfer results when domain shift is large
Classification	1. Cheapest label is required (polyp/non-polyp per frame)	1. Requires considerably more data to converge
Patch-based	1. Increments the size of the training set by obtaining several patches from one single image	1. Slow method in general
Bounding-box	1. Provides enough information for the clinician, focusing their attention on a suspicious area 2. Better computational speed and inter-patch classification coherence 3. Able to converge without pixel-level labels and still predicting location information 4. Bounding-box annotations are cheaper to label than pixel-level ones 5. Preferred type for localization task	1. Does not provide a pixel-level labelling 2. Dense objects might lead to overlapping bouding-boxes
Semantic segmentation	1. Preferred type for segmentation task	1. Accurate borders might be difficult to obtain

location and pathologic diagnosis are also provided. While most polyps present a detected ratio over 80%, only 3 flat polyps (0-IIa in the Paris classification) do not achieve that level, proving the difficulty for detection of this type of polyps.

Characteristics of the polyps are also considered for metrics reporting in the work of Urban et al. [56]. They compare polyp detection by gastroenterologists with and without using the CAD system. In their first study, they found that 9 sessile polyps were missed (i.e. detected with the CAD system and not found without it). Therefore, it might be expected that such system could improve the ADR. Nevertheless, in both studies flat polyps represent a minority: only 2 out of 45 in the first study and 3 out of 73 in the second one. On the other hand, Wang et al. [63] select to separately report results for small (< 0.5 cm), flat, isochromatic polyps, as they are associated with a higher missing rate, finding that the per-image-sensitivity decreased from 0.9438 (95% CI: 0.9380, 0.9496) in the overall dataset to 0.9165 (95% CI: 0.9021, 0.9309). Lastly, Mo et al. [55], although not providing detailed metrics, found that their method behaved differently for sequences showing small polyps, encountering difficulties for their detection.

7. Clinical perspective

The American Society for Gastrointestinal Endoscopy has a set of publications within the Preservation and Incorporation of Valuable endoscopic Innovations (PIVI) initiative to establish thresholds for incorporating innovative technologies into the clinical practice. While there is a PIVI paper related to the in-vivo real-time assessment of diminutive polyps, there is so far no statement regarding the application of CAD systems which could connect the technical development and metrics to clinical performance metrics, such as the ADR.

In this regard, the ADR is considered one of the main indicators for colonoscopy quality, which is influenced both by the endoscopist and the technical factors [117]. Despite presenting a wide variability, ranging from 12.5% to 68.1% in conventional colonoscopy, it has been proven that new technologies such as Endocuff, G-Eye or full-spectrum colonoscopy might help to increase this indicator, reaching values higher than 80% in some of the works systematically reviewed and meta-analyzed by Castaneda et al. [118]. Only two papers in this review [56,63] mention this clinical concept in their technical studies. So, additional efforts should be taken in future works to connect both technical and clinical outcomes to increase the acceptance of CAD systems based on deep learning in the daily clinical practice.

It has been already proven that a second observer, such as an experienced nurse, improves the ADR even in the event of an experienced gastroenterologist performing the colonoscopy [119]. Therefore, and as already Wang et al. [63] suggest, CAD systems might be used as an "extra pair of eyes" to avoid missing subtle lesions. In this regard, they have recently analyzed the influence of their method on the ADR [120]. They have carried out a non-blinded trial, where patients were prospectively randomized into diagnostic colonoscopies with or without

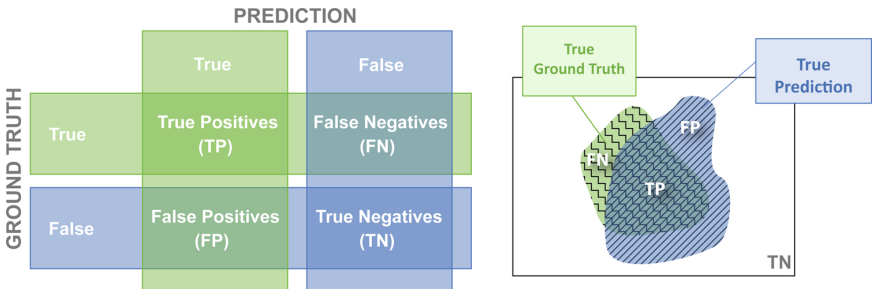


Fig. 10. Confusion matrix. Left side, confusion matrix calculated at frame level. On the right side, elements of the confusion matrix calculated at pixel level by overlapping of two binary masks.

Table 10
Definition of metrics used in the retrieved works.

Metric – alternative names	Calculation
Accuracy	$\mathrm{Acc}_1 = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}$
	$\mathrm{Acc}_2 = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}$
Precision	$\mathrm{Prec} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$
Recall – true positive rate, sensitivity, pre-class accuracy	$\mathrm{Rec} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$
Specificity – pre-class accuracy	$\mathrm{Spec} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}$
F1-score	$F1 = \frac{2\cdot\mathrm{Prec}\cdot\mathrm{Rec}}{\mathrm{Prec}+\mathrm{Rec}}$
F2-score	$F2 = \frac{5\cdot\mathrm{Prec}\cdot\mathrm{Rec}}{4\cdot\mathrm{Prec}+\mathrm{Rec}}$
Matthew correlation coefficient	$\mathrm{MCC} = \frac{\mathrm{TP}\cdot\mathrm{TN}-\mathrm{FP}\cdot\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}$
False positive rate	$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}}$
False positive per frame	$\mathrm{FP/Frames} = \frac{\mathrm{FP}}{\#\,\mathrm{frames}}$
Intersection over Union – Jaccard index	$\mathrm{IoU(PR,\ GR)} = \frac{\lvert\mathrm{PR}\cap\mathrm{GT}\rvert}{\lvert\mathrm{PR}\cup\mathrm{GT}\rvert} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}+\mathrm{FN}} = \frac{\mathrm{Dice}}{2-\mathrm{Dice}}$
Dice coefficient	$\mathrm{Dice(PR,\ GR)} = \frac{2\cdot\lvert\mathrm{PR}\cap\mathrm{GT}\rvert}{\lvert\mathrm{PR}\rvert+\lvert\mathrm{GT}\rvert} = \frac{2\cdot\mathrm{TP}}{2\cdot\mathrm{TP}+\mathrm{FP}+\mathrm{FN}} = \frac{2\cdot\mathrm{IoU}}{1+\mathrm{IoU}}$
Receiver operating characteristic (ROC) Curve	Plot of TPR against FPR
Area under the curve (AUC) – area under the ROC curve (AUROC), ROS curve	$\mathrm{AUC} = 1 - \frac{\mathrm{FPR}-\mathrm{FNR}}{2}$
Free response receiver operating characteristic curve	Plot of TPR against FPR/frame
Temporal coherence	$\mathrm{TC} = \frac{\#\,\text{correctly detected consecutive frame pairs}}{\#\,\text{consecutive frame pairs}}$
Polyp detection rate	$\mathrm{PDR} = \frac{\#\,\text{polyps detected at least in one frame}}{\#\,\text{polyps in the dataset}}$
Mean processing time per frame	Actual detection processing time taken by a method to process a frame and display the detection result
Reaction time	RT = frame of first detection – frame of first appearance
Mean distance	Mean Euclidean distance between polyp centres

TP: True negatives; TN: true negatives; FP: false positives; FN: false negatives; PR: predicted binary mask; GT: ground truth binary mask.

Table 11
Metrics for reporting detection, localization and segmentation tasks.

Task	Metric	Level	Total of works	References
Detection	Accuracy-2	Frame	7	[42,49,51,53–56]
	Precision	Frame	7	[41,42,49,54,55,57,58]
	Recall	Frame	9	[41,42,44,49,51,54,55,57,58]
	Specificity	Frame	5	[42,44,49,51,54]
	F1-score	Frame	3	[54,55,57]
	F2-score	Frame	2	[55,57]
	AUC	Frame	4	[48,51,52,56]
	FPR	Frame	2	[44,49]
	Dice	–	1	[49]
	MCC	Frame	1	[54]
	FROC	Frame	4	[43,45–47]
	ROC	Frame	1	[50]
Localization	Accuracy-2	Frame	2	[64,66]
	Accuracy-2	Patch	1	[54]
	Precision	Frame	7	[55,59–62,65,66]
	Precision	Patch	1	[54]
	Recall	Frame	9	[55,59–66]
	Recall	Patch	1	[54]
	Specificity	Patch	1	[54]
	Specificity	Frame	3	[62–64]
	F1-score	Frame	7	[55,59–62,65,66]
	F1-score	Patch	1	[54]
	F2-score	Frame	5	[55,60–62,65]
	Dice	Pixel	1	[56]
	MCC	Patch	1	[54]
Segmentation	Accuracy-1	Pixel	1	[36]
	Accuracy-2	Pixel	5	[67,68,70,73,74]
	Precision	Pixel	3	[58,69,74]
	Recall	Pixel	4	[58,67,69,74]
	Specificity	Pixel	2	[67,74]
	F1-score	Pixel	2	[69,74]
	IoU	Pixel	6	[36,68–72]
	Dice	Pixel	3	[67–69]

CAD assistance. In this trial, the main outcome was the ADR. Despite not obtaining the highest metrics for the localization task, they prove a significant increment from 20.3% to 29.1% in the ADR, showing the clinical potential of their method. Therefore, the final aim of any technical development should be to prove the benefit in the clinical practice, rather than only raking on the top considering technical metrics.

Another relevant aspect for CAD systems based on deep learning is matching real time constraints to facilitate clinical application in a live procedure. Efforts should therefore be oriented to exploit the use of videos, rather than isolated images, minimizing the processing time to keep it under the restriction of processing 25 or 30 frames per second. Current deep learning algorithms are able to run in nearly real-time speed [111,121–123] and have not been fully tested in real clinical conditions. As far as the authors know, there is only one commercial system for detection assistance based on artificial intelligence [124], but no technical information has been found on its algorithms. We highly advise for the design of real-time clinical essays to analyze the usability and real performance of the algorithms. On the contrary, images in the datasets usually show polyps in a clean, well-centred state. This is though not the situation in which the CAD system is useful. It would be highly interesting to provide "fly-by" explorations as well as difficult polyps to locate, such as partially hidden or located in folders, to mimic situations with higher clinical value. The difference between images available in the dataset and those in the clinical situations might lead to the fact that is not possible to guarantee that success in the dataset will be reproduced in the clinical environment.

Lastly, it is also important to recognize the limitation of CAD systems in the identification of polyps, as they might be missing due to two main situations: (1) they never appear on the visual field, due to an inappropriate bowel preparation, an inappropriate exposure technique or more importantly, because it is in the 20% of the colon surface that is never surveyed [125]; and (2) missed by the gastroenterologist due to a lack of training or short withdrawal time. While CAD systems might

help the gastroenterologist to not miss polyps, they will not counteract the difficulties of poor bowel preparation or exposure technique.

8. Recommendations and future challenges

Polyp detection, localization and segmentation have been boosted in the last years by the application of deep learning strategies. In this review, we have analyzed the datasets, methods and metrics used up to now. Nevertheless, and despite the success of deep learning over handcrafted methods, there are still challenges to be faced by the scientific community in the upcoming years. In this section we aim at pointing out trends and/or give recommendations on future research lines.

Lack of reproducibility has been raising concerns lately as a critical flaw, especially in the field of health as explained by McDermott et al. [126]. In their paper, they pose recommendations for data providers, researchers and journals and conferences, bearing in mind technical, statistical and conceptual replicability as the three main aspects of reproducibility. Many of our recommendations are aligned with their work, particularized to our field of interest.

8.1. Datasets

The collection of images to create a large dataset is one of the challenges that should be addressed by the clinical community, which would eventually help the technical researchers when developing CAD systems. In this regard, it would be useful to collaboratively work under common guidelines, such as the methodology proposed by Sánchez-Peralta et al. [127], that allows for the systematic acquisition and annotation of colonoscopy videos without modifying the clinical routine. This relates to three issues to consider: (1) collecting abnormal cases, (2) the quality of annotations and (3) the variability of acquisition systems. In the first case, since we are dealing with medical images, images from some particular classes might be more difficult to collect, since they are less frequent to be found in the clinical practice despite having even more relevance for the application of the CAD system than other more frequent classes. This is for instance the situation of flat polyps (types 0-IIb and 0-IIc in the Paris classification as showed in Fig. 4). While gastroenterologists find that CAD systems would be more useful to help in their detection, those types are underrepresented in the publicly available databases [96]. This fact might hinder the efficacy of CAD systems. On the other hand, medical datasets must be annotated by clinicians, taking considerable time. Tools such as GTCreator [87] have been designed to ease the process of ground truth creation, allowing for image and text-based annotations, so it would be advisable to use it to share and revise annotations among different clinicians. Strictly speaking, ground truth is not available for detection methods in colonoscopy as annotated datasets are made by expert calls that represent the best knowledge that can be extracted by an expert from the colonoscopy. For that, agreement with expert's praxis can be analyzed by the Cohen's kappa coefficient. In order to minimize this issue, datasets should be independently annotated by different clinicians and the inconsistencies should be handled by an additional clinician. This will allow to create a dataset closer to an actual ground truth. Furthermore, inter-observer variability is a well-known problem in manual segmentation of medical images [128]. It would be advisable that future datasets provide uncertainty maps to reflect the variability of experts opinion on the same image. Lastly, a collaborative dataset would increase the variability of the acquisition systems, therefore strengthening the CAD systems to accurately work regardless of the endoscope manufacturer.

When producing a dataset, it is also important to establish a set of criteria and guidelines, similarly to the challenge rules, so authors can follow them, facilitating the posterior comparison of methods. In this regard, the distribution of images into training, validation and test sets is a minimum where cross validation performed over the different subjects of the dataset or over fixed sets can be employed for deriving the statistical metrics and confidence intervals. It is essential to assure patient independence of the sets, so all images originated from a patient must fall into one of the sets. For testing, it is recommended that the dataset owners establish a bootstrapping methodology [129], what has been successfully used in other domains [130–132], identifying the number of testing images and the corresponding sample size and repetitions, clearly indicating which images must be considered in each iteration. Bootstrapping consists on analyzing different testing subsets and measuring the posterior distribution of the results. Authors are encouraged to report information on the posterior probability of the metrics or, at least, information on their mean and standard deviation, giving a more realistic vision of the method performance.

All currently available datasets exposed in Section 4.1 only contain medical images and videos from actual patients. None of the articles mentions the use of synthetic image datasets, which could be an alternative to increase the number of samples. There are already efforts in this direction, as in the work of Shin et al. [133]. They employ a conditional GAN to synthetize polyp frames from normal colonoscopy images, by using a filtering-based binary image as input, modified to include the position and size of the polyp. Even though the generated images are qualitatively realistic, they result on deterministic polyps without many variations on colour and texture. Hence, the potential of synthetic image can be further explored and exploited.

8.2. Metrics

Another element to define for fair comparison is the set of metrics to be used for reporting. As seen in Section 6, there is a lack of criteria among the authors to select the most convenient one. In this choice, there are two aspects that play a role. On one side, the type of ground truth available and on the other side, the task to be accomplished. While the former limits the available information (label per image/frame or binary mask), the latter relates to the information worthy to measure. For methods aiming at detecting and locating polyps, it would be advisable to calculate the F1-score at frame level, as it gives a balanced measure between missing polyps (or false negatives – FN) and false alarms (or false positives – FP) [24]. In order to trace easier parallels with the literature of computer vision when analyzing results, we would also recommend including mAP to perform a global technical evaluation of the algorithm, as commented in Section 6. However, mAP analysis should be complemented with a more detailed analysis to validate the real clinical performance of the model. As for segmentation methods, despite being useful, metrics based on elements of the confusion matrix at pixel-level do not detect whether the two masks are similar in shape [114], so it would be useful to complement them with distance metrics that are valid for small segments (such as Hausdorff distance or Mahalanobis distance). In this regard, it is also important to mention the recommendation to calculate agreement measurements with the experts. Specially in the case of detection methods, it would be highly desirable to compute inter-rater measurements, selecting the most suitable method depending on the type of variable (categorical or continuous) and the number of observers [134].

Clinically speaking and agreeing with the concern raised by Robinson et al. [135], we are of the opinion that reporting metrics per patient or per polyp might be more convenient than averaging results all over the test set, as long as the database provides information to identify which polyp and/or patient originate each frame. Polyps and/or patients might have an unequal presence in the database, for example polyp A and B having 8 and 2 frames, respectively, in the test set. If a detection method has 80% accuracy, it might be that all detected frames corresponds to polyp A. If metrics are averaged all over the test set, this situation cannot be identified, but if metrics are provided per polyp (and later averaged), accuracy would be 50%. This way, it would be possible to identify the cases in which the method presents flaws to further work on.

8.3. Data augmentation

In terms of data augmentation, it has been shown that transformations are selected based on subjective criteria upon the researcher experience and that there is no general strategy as the differences in Table 5 show. Efforts are therefore now focused on the identification of the most convenient transformations and ranges in a more objective way. Initiatives to find the most convenient data augmentation policies, such as Smart Augmentation [98], AutoAugment [136] or the use of a Bayesian data augmentation approach [137] have been mostly developed for classification of natural images. Thus, the application of these methodologies to colonoscopy images might boost performance of methods and is therefore worth research.

Besides, it would be also interesting that data augmentation transformations would address particularities of the colonoscopy images, such as illumination effects (specular lights and lack of uniformity); sensor acquisition effects (colour phantoms); image interlacing; sharpening (to improve the quality of the visualized image but increasing the image noise at the same time); information overlay or the presence of the black mask [138]. As these effects might negatively affect the CAD system performance, their inclusion in the training dataset could lead to the model invariance when they are present.

8.4. Network design

The utility of CNNs for polyp detection, localization and segmentation have been already widely explored using supervised learning. In the future, semi-supervised or unsupervised training should be further exploited, relying on smaller datasets which would be eventually easier to compile. On the other hand, LTSMs or RNNs, with small presence in the current review, will be more widely employed in the field of CRC detection. The capability of recurrent networks to model temporal relationships can help creating models tackling temporal information into account for colonoscopy videos. On the other hand, GANs can be employed in the future to generate synthetic data from polyp models to increase the variability of the dataset or to be able to over-express difficult to detect lesions into existing datasets.

Besides, networks have to be designed for taking advantage for the new very high-resolution colonoscopy devices. This can help detecting micro-patterns that can be missed by veteran gastroenterologists, specially from small or flat polyps. This implies challenges on the definition of a network capable of real-time inspection capabilities for very high resolution images.

Reproducibility of deep learning methods comes with associated difficulties. It should be clearly stated in the papers the following information:

- Dataset and distribution of training, validation and test subsets.
- Data pre-processing, if any.
- Model training, including learning parameters such as learning rate, early stopping or weight initialization. The use of seeds when applicable.
- Loss function, as it highly impacts the model performance.
- Hardware and software details, as software packages are continuously updated, and some models might require exceptional hardware conditions.

Randomization of parameters hampers reproducibility, e.g. when images are randomly transformed on the flow. Although ideally all particular seeds and values should be reported, or the code made available, so other researchers could reproduce the experiment, releasing at least the trained models could be an intermediate solution.

Some efforts on reproducibility have also been taken from major conferences organizations such as NeurIPS, with the request of a reproducibility checklist [139]. We strongly advise its use in future publications related to CAD systems based on deep learning.

9. Conclusions

CRC is one of the major causes of death by cancer worldwide. Early detection of precursor lesions has been proved to minimize its incidence, so screening programs are essential. Colonoscopy is a gold standard technique for the detection and treatment of polyps and adenomas. CAD systems might help endoscopists to identify lesions and minimize the ADR.

In the current work, we provided a systematic and comprehensive review of 35 works for detection, localization and segmentation of polyps using deep learning approaches since 2015. We further analyzed seven currently available public databases of colonoscopy images as well as the most common metrics used for reporting. Retrieved methods have been classified according to the approach they follow in a primary (end-to-end vs hybrid methods) and secondary (feature extractor, classification, patch-based, bounding-box and semantic segmentation) classifications. Although there is no common dataset or framework for easy and direct comparison of methods, some trends, advantages and disadvantages have been identified and discussed. Lastly, recommendations and future challenges have been identified.

Despite the great success of deep learning approaches, clinical validation and application is still a must. The creation of larger, more assorted, public datasets; new algorithms requesting less training samples and the creation of a common validation framework will maintain the upwards tendency and will end in the clinical application of CAD systems to assist gastroenterologists to increase the ADR and early detect CRC.

Conflict of interest

The authors declare that there is no conflict of interest.

Acknowledgments

This work was partially supported by PICCOLO project. This project has received funding from the European Union's Horizon2020 Research and Innovation Programme under grant agreement No. 732111. The sole responsibility of this publication lies with the author. The European Union is not responsible for any use that may be made of the information contained therein. The authors would also like to thank Dr. Federico Soria for his support on this manuscript and Dr. José Carlos Marín, from Hospital 12 de Octubre, and Dr. Ángel Calderón and Dr. Francisco Polo, from Hospital de Basurto, for the images in Fig. 4.

References

[1] World Health Organization. World cancer report 2014. 2014.
[2] International Agency for Research on Cancer. Colorectal cancer factsheet. Tech. rep. 2018http://gco.iarc.fr/today/data/factsheets/cancers/10_8_9-Colorectum-fact-sheet.pdf.
[3] Siegel RL, Miller KD, Jemal A. Cancer statistics, 2015. CA Cancer J Clin 2015;65(1):29. https://doi.org/10.3322/caac.21254.
[4] Siegel RL, Miller KD, Jemal A. Cancer statistics, 2019. CA Cancer J Clin 2019;69:7–34. https://doi.org/10.3322/caac.21387.
[5] Ferlay J, Colombet M, Soerjomataram I, Dyba T, Randi G, Bettio M, et al. Cancer incidence and mortality patterns in Europe: estimates for 40 countries and 25 major cancers in 2018. Eur J Cancer 2018;103:356–87. https://doi.org/10.1016/j.ejca.2018.07.005.
[6] Wiegering A, Ackermann S, Riegel J, Dietz UA, Götze O, Germer CT, et al. Improved survival of patients with colon cancer detected by screening colonoscopy. Int J Colorectal Dis 2016;31(5):1039–45. https://doi.org/10.1007/s00384-015-2501-6.
[7] Williams CB. Insertion technique. In: Waye JD, Rex DK, Williams CB, editors. Colonoscopy. Principles and practice Blackwell Publishing Ltd; 2009. https://doi.org/10.1002/9781444316902.ch40.
[8] Berros Fombella JP, Aguilar Huergo S, García Teijido P. Enfermedades premalignas. In: Sociedad Espa nola de Oncología Médicaeditor. Manual SEOM de prevención y diagnóstico precoz del cáncer. 2017https://seom.org/seomcms/images/stories/recursos/Manual_SEOM_Prevencion_2017.pdf.
[9] Kamiński MF, Hassan C, Bisschops R, Pohl J, Pellisé M, Dekker E, et al. Advanced imaging for detection and differentiation of colorectal neoplasia: European Society

of Gastrointestinal Endoscopy (ESGE) Guideline. Endoscopy 2014;46:435–49. https://doi.org/10.1055/s-0034-1365348.
[10] Müller MF, Ibrahim AEK, Arends MJ. Molecular pathological classification of colorectal cancer. Virchows Arch 2016;469(2):125–34. https://doi.org/10.1007/s00428-016-1956-3.
[11] Wieszczy P, Regula J, Kaminski M. Adenoma detection rate and risk of colorectal cancer. Best Pract Res Clin Gastroenterol 2017;31:441–6. https://doi.org/10.1016/j.bpg.2017.07.002.
[12] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521(7553):436–44. https://doi.org/10.1038/nature14539.
[13] Ker J, Wang L, Rao J, Lim T. Deep learning applications in medical image analysis. IEEE Access 2018;6:9375–89. https://doi.org/10.1109/ACCESS.2017.2788044.
[14] Hannun AY, Rajpurkar P, Haghpanahi M, Tison GH, Bourn C, Turakhia MP, et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nat Med 2019;25:65–9. https://doi.org/10.1038/s41591-018-0268-3.
[15] Wei JW, Tafe LJ, Linnik YA, Vaickus LJ, Tomita N, Hassanpour S. Pathologist-level classification of histologic patterns on resected lung adenocarcinoma slides with deep neural networks. Sci Rep 2019;9(1):3358. https://doi.org/10.1038/s41598-019-40041-7.
[16] Yu S, Xiao D, Frost S, Kanagasingam Y. Robust optic disc and cup segmentation with deep learning for glaucoma detection. Comput Med Imaging Graph 2019;74:61–71. https://doi.org/10.1016/j.compmedimag.2019.02.005.
[17] Byrne MF, Shahidi N, Rex DK. Will Computer-aided detection and diagnosis revolutionize colonoscopy? Gastroenterology 2017;153(6). https://doi.org/10.1053/j.gastro.2017.10.026. 1460–1464.e1.
[18] Iwahori Y, Hagi H, Usami H, Woodham RJ, Wang A, Bhuyan MK, et al. Automatic polyp detection from endoscope image using likelihood map based on edge information. Proceedings of the 6th international conference on pattern recognition applications and methods – volume 1: ICPRAM 2017:402–9. https://doi.org/10.5220/0006189704020409.
[19] Iakovidis DK, Maroulis DE, Karkanis SA. An intelligent system for automatic detection of gastrointestinal adenomas in video endoscopy. Comput Biol Med 2006;36(10):1084–103. https://doi.org/10.1016/j.compbiomed.2005.09.008.
[20] Ameling S, Wirth S, Paulus D, Lacey G, Vilarino F. Texture-based polyp detection in colonoscopy. Bildverarbeitung für die Medizin 2009 2009:346–50. https://doi.org/10.1007/978-3-540-93860-6_70.
[21] Bernal J, Sánchez FJ, Fernández-Esparrach G, Gil D, Rodríguez de Miguel C, Vilari no F. WM-DOVA maps for accurate polyp highlighting in colonoscopy: validation vs. saliency maps from physicians. Comput Med Imaging Graph 2015;43:99–111. https://doi.org/10.1016/j.compmedimag.2015.02.007.
[22] Bernal J, Sánchez FJ, Vilariño F. Towards automatic polyp detection with a polyp appearance model. Pattern Recognit 2012;45:3166–82. https://doi.org/10.1016/j.patcog.2012.03.002.
[23] Fu JJ, Yu Y-W, Lin H-M, Chai J-W, Chen CC-C. Feature extraction and pattern classification of colorectal polyps in colonoscopic imaging. Comput Med Imaging Graph 2014;38(4):267–75. https://doi.org/10.1016/j.compmedimag.2013.12.009.
[24] Bernal J, Tajbakhsh N, Sánchez FJ, Matuszewski BJ, Chen H, Yu L, et al. Comparative validation of polyp detection methods in video colonoscopy: results from the MICCAI 2015 endoscopic vision challenge. IEEE Trans Med Imaging 2017;36(6):1231–49. https://doi.org/10.1109/TMI.2017.2664042.
[25] Biswas M, Kuppili V, Saba L, Edla DR, Suri HS, Cuadrado-godia E, et al. State-of-the-art review on deep learning in medical imaging. Front Biosci 2019;24:380–406. https://doi.org/10.2741/4725.
[26] Kim J, Hong J, Park H. Prospects of deep learning for medical imaging. Precis Future Med 2018;2(2):37–52. https://doi.org/10.23838/pfm.2018.00030.
[27] Anwar SM, Majid M, Qayyum A, Awais M, Alnowami M, Khan MK. Medical image analysis using convolutional neural networks: a review. J Med Syst 2018;42(11):1–13. https://doi.org/10.1007/s10916-018-1088-1.
[28] Shen D, Wu G, Suk H-I. Deep learning in medical image analysis. Annu Rev Biomed Eng 2017;19(1):221–48. https://doi.org/10.1146/annurev-bioeng-071516-044442.
[29] Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Med Image Anal 2017;42:60–88. https://doi.org/10.1016/j.media.2017.07.005.
[30] Suzuki K. Overview of deep learning in medical imaging. Radiol Phys Technol 2017;10(3):257–73. https://doi.org/10.1007/s12194-017-0406-5.
[31] Greenspan H, van Ginneken B, Summers RM. Deep learning in medical imaging: overview and future promise of an exciting new technique. IEEE Trans Med Imaging 2016;35(5):1153–9. https://doi.org/10.1109/TMI.2016.2553401.
[32] Prasath VBS. Polyp detection and segmentation from video capsule endoscopy: a review. J Imaging 2017;3(1):1–15. https://doi.org/10.3390/jimaging3010001.
[33] Ahmad OF, Soares AS, Mazomenos E, Brandao P, Vega R, Seward E, et al. Artificial intelligence and computer-aided diagnosis in colonoscopy: current evidence and future directions. Lancet Gastroenterol Hepatol 2019;4(1):71–80. https://doi.org/10.1016/S2468-1253(18)30282-6.
[34] Danelakis A, Theoharis T, Verganelakis DA. Survey of automated multiple sclerosis lesion segmentation techniques on magnetic resonance imaging. Comput Med Imaging Graph 2018;70:83–100. https://doi.org/10.1016/j.compmedimag.2018.10.002.
[35] EndoVis Grand Challenge. EndoVis grand challenge. 2019https://endovis.grand-challenge.org/endoscopic_vision_challenge/.
[36] Vázquez D, Bernal J, Sánchez FJ, Fernández-Esparrach G, López AM, Romero A, et al. A benchmark for endoluminal scene segmentation of colonoscopy images. Journal of Healthcare Engineering 2017;2017:4037190. https://doi.org/10.1155/2017/4037190.
[37] Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JP, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. J Clin Epidemiol 2009;62(10):e1–34. https://doi.org/10.1016/j.jclinepi.2009.06.006.
[38] Pogorelov K, Lange TD, Randel KR, Dang-Nguyen D-T, Johansen D, Riegler M. A holistic multimedia system for gastrointestinal tract disease detection. Proceedings of the 8th ACM on multimedia systems conference (MMSys'17) 2017:112–23. https://doi.org/10.1145/3083187.3083189.
[39] Park SY, Sargent D. Colonoscopic polyp detection using convolutional neural networks. Medical imaging 2016: computer-aided diagnosis, vol. 9785 2016. https://doi.org/10.1117/12.2217148.
[40] Ribeiro E, Uhl A, Wimmer G, Häfner M. Exploring deep learning and transfer learning for colonic polyp classification. Comput Math Methods Med 2016;2016:6584725. https://doi.org/10.1155/2016/6584725.
[41] Taha B, Dias J, Wergui N. Convolutional neural network as a feature extractor for automatic polyp detection. 24th IEEE international conference on image processing, ICIP 2017 2017:2060–4. https://doi.org/10.1109/ICIP. 2017.8296644.
[42] Shin Y, Balasingham I, Member S. Comparison of hand-craft feature based svm and cnn based deep learning framework for automatic polyp classification. 2017 39th annual international conference of the IEEE engineering in medicine and biology society (EMBC) 2017:3277–80. https://doi.org/10.1109/EMBC.2017.8037556.
[43] Tajbakhsh N, Shin JY, Gurudu SR, Hurst RT, Kendall CB, Gotway MB, et al. On the necessity of fine-tuned convolutional neural networks for medical imaging. Deep learning and convolutional neural networks for medical image computing: precision medicine, high performance and large-scale datasets 2017:181–93. https://doi.org/10.1007/978-3-319-42999-1_11.
[44] Yuan Z, Izadyyazdanabadi M, Mokkapati D, Panvalkar R, Shin JY, Tajbakhsh N, et al. Automatic polyp detection in colonoscopy videos. Prog Biomed Opt Imaging – Proc SPIE 2017;10133:1–10. https://doi.org/10.1117/12.2254671.
[45] Tajbakhsh N, Shin JY, Gurudu SR, Hurst RT, Kendall CB, Gotway MB, et al. Convolutional neural networks for medical image analysis: full training or fine tuning? IEEE Trans Med Imaging 2016;35(5):1299–312. https://doi.org/10.1109/TMI.2016.2535302.
[46] Tajbakhsh N, Gurudu SR, Liang J. Automatic polyp detection in colonoscopy videos using an ensemble of convolutional neural networks. 2015 IEEE 12th international symposium on biomedical imaging 2015:79–83. https://doi.org/10.1109/ISBI.2015.7163821.
[47] Tajbakhsh N, Gurudu SR, Liang J. A comprehensive computer-aided polyp detection system for colonoscopy videos. 24th International Conference on Information Processing in Medical Imaging, IPMI 2015 2015:327–38. https://doi.org/10.1007/978-3-319-19992-4_25.
[48] Axyonov S, Zamyatin A, Liang J, Kostin K. Advanced pattern recognition and deep learning for colon polyp detection. Distributed computer and communication networks: control, computation, communications (DCCN-2016) 2016:27–34http://vital.lib.tsu.ru/vital/access/manager/Repository/vtls:000616307.
[49] Akbari M, Mohrekesh M, Rafiei S, Reza Soroushmehr SM, Karimi N, Samavi S, et al. Classification of informative frames in colonoscopy videos using convolutional neural networks with binarized weights. 2018 40th annual international conference of the IEEE engineering in medicine and biology society (EMBC), vol. 2018 2018:65–8. https://doi.org/10.1109/EMBC.2018.8512226. arXiv:1802.01387.
[50] Aksenov S, Kostin K, Ivanova A, Liang J, Zamyatin A. An ensemble of convolutional neural networks for the use in video endoscopy. Sovrem Tehnol Med 2018;10(2):7–17. https://doi.org/10.17691/stm2018.10.2.01.
[51] Itoh H, Roth HR, Lu L, Oda M, Misawa M, Mori Y, et al. Towards automated colonoscopy diagnosis: binary polyp size estimation via unsupervised depth learning. medical image computing and computer-assisted intervention. MICCAI 2018. Lecture notes in computer science, vol 11071 2018:611–9. https://doi.org/10.1007/978-3-030-00934-2_68.
[52] Misawa M, Kudo SE, Mori Y, Cho T, Kataoka S, Yamauchi A, et al. Artificial intelligence-assisted polyp detection for colonoscopy: initial experience. Gastroenterology 2018;154(8):2027–9. https://doi.org/10.1053/j.gastro.2018.04.003.
[53] Murthy VN, Singh V, Sun S, Bhattacharya S, Chen T, Comaniciu D. Cascaded deep decision networks for classification of endoscopic images. Proceedings volume 10133, medical imaging 2017: image processing 2017. https://doi.org/10.1117/12.2254333. pp. 10133-1–10133-15.
[54] Pogorelov K, Ostroukhova O, Jeppsson M, Espeland HN, Griwodz C, De Lange T, et al. Deep learning and hand-crafted feature based approaches for polyp detection in medical videos. 2018. p. 381–6. https://doi.org/10.1109/BHI.2018.8333444.
[55] Mo X, Tao K, Wang Q, Wang G. An efficient approach for polyps detection in endoscopic videos based on faster R-CNN. 2018 24th international conference on pattern recognition (ICPR) 2018:3929–34https://arxiv.org/abs/1809.01263.
[56] Urban G, Tripathi P, Alkayali T, Mittal M, Jalali F, Karnes W, et al. Accuracy in screening colonoscopy. Gastroenterology 2018;155(4). https://doi.org/10.1053/j.gastro.2018.06.037. 1069–1078.e8.
[57] Mohammed A, Yildirim S, Farup I, Pedersen M, Hovde Ø. Y-Net: a deep convolutional neural network for polyp detection. 29th British machine vision conference (BMCV2018) 2018http://bmvc2018.org/contents/papers/0487.pdf.
[58] Brandao P, Mazomenos E, Ciuti G, Caliò R, Bianchi F, Menciassi A, et al. Fully convolutional neural networks for polyp segmentation in colonoscopy. Medical imaging 2017: computer-aided diagnosis, vol. 10134 2017:101340F. https://doi.org/10.1117/12.2254361.
[59] Park S, Lee M, Kwak N. Polyp detection in colonoscopy videos using deeply-

learned hierarchical features. 2015http://mipal.snu.ac.kr/images/0/0b/Polyp_short_report.pdf.
[60] Shin Y, Qadir HA, Aabakken L, Bergsland J, Balasingham I. Automatic colon polyp detection using region based deep CNN and post learning approaches. IEEE Access 2018;6:40950–62. https://doi.org/10.1109/ACCESS.2018.2856402.
[61] Yu L, Chen H, Dou Q, Qin J, Heng PA. Integrating online and offline three-dimensional deep learning for automated polyp detection in colonoscopy videos. IEEE J Biomed Health Inform 2017;21(1):65–75. https://doi.org/10.1109/JBHI.2016.2637004.
[62] Zhang R, Zheng Y, Poon CC, Shen D, Lau JY. Polyp detection during colonoscopy using a regression-based convolutional neural network with a tracker. Pattern Recognit 2018;83:209–19. https://doi.org/10.1016/j.patcog.2018.05.026.
[63] Wang P, Xiao X, Glissen Brown JR, Berzin TM, Tu M, Xiong F, et al. Development and validation of a deep-learning algorithm for the detection of polyps during colonoscopy. Nat Biomed Eng 2018;2:741–8. https://doi.org/10.1038/s41551-018-0301-3.
[64] Billah M, Waheed S, Rahman MM. An automatic gastrointestinal polyp detection system in video endoscopy using fusion of color wavelet and convolutional neural network features. Int J Biomed Imaging 2017;2017:1–10. https://doi.org/10.1155/2017/9545920.
[65] Zheng Y, Zhang R, Yu R, Jiang Y, Mak TWC, Wong SH, et al. Localisation of colorectal polyps by convolutional neural network features learnt from white light and narrow band endoscopic images of multiple databases. 2018 40th annual international conference of the IEEE engineering in medicine and biology society (EMBC) 2018:4142–5. https://doi.org/10.1109/EMBC.2018.8513337.
[66] Pogorelov K, Riegler M, Eskeland SL, de Lange T, Johansen D, Griwodz C, et al. Efficient disease detection in gastrointestinal videos – global features versus neural networks. Multimed Tools Appl 2017;76(21):22493–525. https://doi.org/10.1007/s11042-017-4989-y.
[67] Zhang L, Dolwani S, Ye X. Automated polyp segmentation in colonoscopy frames using fully convolutional neural network and textons. In: Valdes Hernandez M, González-Castro V, editors. Medical image understanding and analysis. MIUA 2017. Communications in computer and information science, vol. 723 Springer; 2017. p. 707–17. https://doi.org/10.1007/978-3-319-60964-5_62.
[68] Nguyen Q, Lee S-W. Colorectal segmentation using multiple encoder-decoder network in colonoscopy images. 2018 IEEE first international conference on artificial intelligence and knowledge engineering (AIKE) 2018:208–11. https://doi.org/10.1109/AIKE.2018.00048.
[69] Wichakam I, Panboonyuen T, Udomcharoenchaikit C. Real-time polyps segmentation for colonoscopy video frames using compressed fully convolutional network. MultiMedia modeling. MMM 2018. Lecture notes in computer science, vol 10704 2018:393–404. https://doi.org/10.1007/978-3-319-73603-7_32.
[70] Wickstrøm K, Kampffmeyer M, Jenssen R. Uncertainty modeling and interpretability in convolutional neural networks for polyp segmentation. 2018 IEEE international workshop on machine learning for signal processing 2018. https://doi.org/10.1109/MLSP.2018.8516998.
[71] Xiao W-T, Chang L-J, Liu W-M. Semantic segmentation of colorectal polyps with DeepLab and LSTM networks. 2018 IEEE international conference on consumer electronics-Taiwan (ICCE-TW) semantic 2018:115–20. https://doi.org/10.1109/ICCE-China.2018.8448568.
[72] Zhou Z, Siddiquee MMR, Tajbakhsh N, Liang J. UNet++: a nested U-Net architecture for medical image segmentation. Deep learning in medical image analysis and multimodal learning for clinical decision support. DLMIA 2018, ML-CDS 2018. Lecture notes in computer science, vol. 11045 2018. https://doi.org/10.1007/978-3-030-00889-5_1.
[73] Bardhi O, Sierra-Sosa D, Garcia-Zapirain B, Elmaghraby A. Automatic colon polyp detection using convolutional encoder-decoder model. 2017 IEEE international symposium on signal processing and information technology (ISSPIT) 2017:445–8. https://doi.org/10.1109/ISSPIT.2017.8388684.
[74] Li Q, Yang G, Chen Z, Huang B, Chen L, Xu D, et al. Colorectal polyp segmentation using a fully convolutional neural network. 2017 10th international congress on image and signal processing, biomedical engineering and informatics (CISP-BMEI 2017) 2017:1–5. https://doi.org/10.1109/CISP-BMEI.2017.8301980.
[75] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Advances in neural information processing systems 25 (NIPS2012) 2012. https://doi.org/10.1145/3065386.
[76] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2015http://arxiv.org/abs/1409.1556.
[77] Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions. 2015 IEEE conference on computer vision and pattern recognition (CVPR) 2015. https://doi.org/10.1109/CVPR.2015.7298594.
[78] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. 2016 IEEE conference on computer vision and pattern recognition (CVPR) 2016. https://doi.org/10.1109/CVPR.2016.90.
[79] Badrinarayanan V, Handa A, Cipolla R. SegNet: a deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling. IEEE Trans Pattern Anal Mach Intell 2017;39(12):2481–95. https://doi.org/10.1103/PhysRevX.5.041024.
[80] Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells WM, Frangi AF, editors. Medical image computing and computer-assisted intervention. MICCAI 2015. Lecture notes in computer science, vol. 9351 Springer; 2015. p. 234–41. https://doi.org/10.1007/978-3-319-24574-4_28.
[81] Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. ImageNet: a large-scale hierarchical image database. 2009 IEEE conference on computer vision and pattern recognition 2009. https://doi.org/10.1109/CVPR.2009.5206848.
[82] Stanford Vision Lab. Stanford University. Princeton University, ImageNet dataset. http://www.image-net.org/.
[83] Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, et al. Microsoft COCO: common objects in context. Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics) 8693 LNCS (Part 5) 2014:740–55. https://doi.org/10.1007/978-3-319-10602-1_48.
[84] COCO Consortium, MSCoco dataset. http://cocodataset.org/#home.
[85] Everingham M, Eslami SMA, Van Gool L, Williams CKI, Winn J, Zisserman A. The pascal visual object classes challenge: a retrospective. Int J Comput Vis 2014;111(1):98–136. https://doi.org/10.1007/s11263-014-0733-5.
[86] PASCAL VOC project. PascalVOC dataset. http://host.robots.ox.ac.uk/pascal/VOC/index.html.
[87] Bernal J, Histace A, Masana M, Angermann Q, Sánchez-Montes C, de Miguel CR, et al. GTCreator: a flexible annotation tool for image-based datasets. Int J Comput Assist Radiol Surg 2019;14(2):191–201. https://doi.org/10.1007/s11548-018-1864-x.
[88] Tajbakhsh N, Gurudu SR, Liang J. Automated polyp detection in colonoscopy videos using shape and context information. IEEE Trans Med Imaging 2016;35(2):630–44. https://doi.org/10.1109/TMI.2015.2487997.
[89] Angermann Q, Bernal J, Sánchez-Montes C, Hammami M, Fernández-Esparrach G, Dray X, et al. Towards real-time polyp detection in colonoscopy videos: adapting still frame-based methodologies for video sequences analysis. Computer assisted and robotic endoscopy and clinical image-based procedures 2017:29–41. https://doi.org/10.1007/978-3-319-67543-5.
[90] Pogorelov K, Randel KR, Griwodz C, Eskeland SL, de Lange T, Johansen D, et al. Kvasir: a multi-class image dataset for computer aided gastrointestinal disease detection. Proceedings of the 8th ACM multimedia systems conference (MMSys'17) 2017:164–9. https://doi.org/10.1145/3083187.3083212.
[91] Consortium for Open Medical Image Computing, Sub-Challenge Gastrointestinal Image ANAlysis (GIANA); 2017. https://giana.grand-challenge.org/Home/.
[92] Pogorelov K, Randel KR, de Lange T, Eskeland SL, Griwodz C, Johansen D, et al. Nerthus: a bowel preparation quality video dataset. Proceedings of the 8th ACM on multimedia systems conference (MMSys'17) 2017:170–4. https://doi.org/10.1145/3083187.3083216.
[93] Mesejo P, Pizarro D, Abergel A, Rouquette O, Beorchia S, Poincloux L, et al. Computer-aided classification of gastrointestinal lesions in regular colonoscopy. IEEE Trans Med Imaging 2016;35(9):2051–63. https://doi.org/10.1109/TMI.2016.2547947.
[94] Participants in the Paris Workshop. The Paris endoscopic classification of superficial neoplastic lesions: esophagus, stomach, and colon. Gastrointest Endosc 2003;58(6):S3–43. https://doi.org/10.1016/S0016-5107(03)02159-X.
[95] Endoscopic Classification Review Group. Update on the Paris classification of superficial neoplastic lesions in the digestive tract. Endoscopy 2005;37(6):570–8. https://doi.org/10.1055/s-2005-861352.
[96] Sánchez-Peralta LF, Sánchez-Margallo FM, Bote Chacón J, Soria Gálvez F, López-Saratxaga C, Picón Ruiz A, et al. Is it necessary to improve the colorectal polyps databases for detection CAD systems based on deep learning? Br J Surg 2018;105(S2):5–14.
[97] Eaton-Rosen Z, Bragman F, Ourselin S, Cardoso MJ. Improving data augmentation for medical image segmentation. 1st conference on medical imaging with deep learning (MIDL 2018) 2018. . https://openreview.net/pdf?id=rkBBChjiG.
[98] Lemley J, Bazrafkan S, Corcoran P. Smart augmentation learning an optimal data augmentation strategy. IEEE Access 2017;5:5858–69. https://doi.org/10.1109/ACCESS.2017.2696121.
[99] Goodfellow I, Bengio Y, Courville A. Deep learning. MIT Press; 2016http://www.deeplearningbook.org/.
[100] Pouyanfar S, Sadiq S, Yan Y, Tian H, Tao Y, Reyes MP, et al. A survey on deep learning: algorithms, techniques and applications. ACM Comput Surv 2018;51(5). https://doi.org/10.1145/3234150.
[101] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. 2015 IEEE conference on computer vision and pattern recognition (CVPR) 2015:3431–40. https://doi.org/10.1109/CVPR.2015.7298965.
[102] Jegou S, Drozdzal M, Vazquez D, Romero A, Bengio Y. The one hundred layers tiramisu: fully convolutional densenets for semantic segmentation. 2017 IEEE conference on computer vision and pattern recognition workshops (CVPRW) 2017:1175–83. https://doi.org/10.1109/CVPRW.2017.156.
[103] Marron JS, Todd MJ, Ahn J. Distance-weighted discrimination. J Am Stat Assoc 2007;102(480):1267–71. https://doi.org/10.2307/27639976.
[104] Heberle H, Meirelles VG, da Silva FR, Telles GP, Minghim R. InteractiVenn: a web-based tool for the analysis of sets through Venn diagrams. BMC Bioinform 2015;16(1):1–7. https://doi.org/10.1186/s12859-015-0611-3.
[105] Moriya T, Roth HR, Nakamura S, Oda H, Nagara K, Oda M, et al. Unsupervised segmentation of 3D medical images based on clustering and deep representation learning. Proceedings volume 10578, medical imaging 2018: biomedical applications in molecular, structural, and functional imaging 2018. https://doi.org/10.1117/12.2293414.
[106] Medela A, Picon A, Saratxaga CL, Belar O, Cabezón V, Cicchi R, et al. Few shot learning in histopathological images: reducing the need of labeled data on biological datasets. 2019 IEEE 16th international symposium on biomedical imaging (ISBI) 2019.
[107] Medela A, Picon A. Constellation loss: improving the efficiency of deep metric learning loss functions for optimal embedding. 2019https://arxiv.org/abs/1905.10675.
[108] Tran D, Bourdev L, Fergus R, Torresani L, Paluri M. Learning spatiotemporal features with 3D convolutional networks. Proceedings of the IEEE international conference on computer vision 2015 Inter 2015:4489–97. https://doi.org/10.

1109/ICCV.2015.510.
[109] Chollet F. Xception: deep learning with depthwise separable convolutions. 2017 IEEE conference on computer vision and pattern recognition (CVPR) 2017:1800–7. https://doi.org/10.1109/CVPR.2017.195.
[110] Szegedy C, Ioffe S, Vanhoucke V, Alemi AA. Inception-v4, Inception-ResNet and the impact of residual connections on learning. Thirty-first AAAI conference on artificial intelligence (AAAI-17) 2017:4278–84https://www.aaai.org/ocs/index.php/AAAI/AAAI17/paper/download/14806/14311.
[111] Redmon J, Farhadi A. YOLOv3: an incremental improvement. 2018http://arxiv.org/abs/1804.02767.
[112] Son J, Park SJ, Jung K-H. Retinal vessel segmentation in fundoscopic images with generative adversarial networks. 2017http://arxiv.org/abs/1706.09318.
[113] Sudre CH, Li W, Vercauteren T, Ourselin S, Jorge Cardoso M, et al. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In: Cardoso MJ, Arbel T, Carneiro G, Syeda-Mahmood T, Tavares JMR, Moradi M, editors. Deep learning in medical image analysis and multimodal learning for clinical decision support. DLMIA 2017, ML-CDS 2017. Lecture notes in computer science, vol. 10553 Springer; 2017. p. 240–8. https://doi.org/10.1007/978-3-319-67558-9_28.
[114] Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. BMC Med Imaging 2015;15:29. https://doi.org/10.1186/s12880-015-0068-x.
[115] Everingham M, Van Gool L, Williams CK, Winn J, Zisserman A. The pascal visual object classes (VOC) challenge. Int J Comput Vis 2010;88(2):303–38. https://doi.org/10.1007/s11263-009-0275-4.
[116] Zhang X, Chen F, Yu T, An J, Huang Z, Liu J, et al. Real-time gastric polyp detection using convolutional neural networks. PLoS ONE 2019;14(3):1–16. https://doi.org/10.1371/journal.pone.0214133.t005.
[117] Rex DK. Polyp detection at colonoscopy: endoscopist and technical factors. Best Pract Res: Clin Gastroenterol 2017;31(4):425–33. https://doi.org/10.1016/j.bpg.2017.05.010.
[118] Castaneda D, Popov VB, Verheyen E, Wander P, Gross SA. New technologies improve adenoma detection rate, adenoma miss rate, and polyp detection rate: a systematic review and meta-analysis. Gastrointest Endosc 2018;88(2):209–22. https://doi.org/10.1016/j.gie.2018.03.022.
[119] Wang W, Xu L, Bao Z, Sun L, Hu C, Zhou F, et al. Differences with experienced nurse assistance during colonoscopy in detecting polyp and adenoma: a randomized clinical trial. Int J Colorectal Dis 2018;33(5):561–6. https://doi.org/10.1007/s00384-018-3003-0.
[120] Wang P, Berzin TM, Glissen Brown JR, Bharadwaj S, Becq A, Xiao X, et al. Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: a prospective randomised controlled study. Gut 2019:1–7. https://doi.org/10.1136/gutjnl-2018-317500.
[121] Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C-Y, et al. Ssd: single shot multibox detector. In: Leibe B, Matas J, Sebe N, Welling M, editors. Computer vision – ECCV 2016. Springer International Publishing; 2016. p. 21–37.
[122] Ren S, He K, Girshick R, Sun J. Faster r-cnn: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell 2017;39(6):1137–49. https://doi.org/10.1109/TPAMI.2016.2577031.
[123] He K, Gkioxari G, Dollár P, Girshick R. Mask r-cnn. 2017 IEEE international conference on computer vision (ICCV) 2017:2980–8. https://doi.org/10.1109/ICCV.2017.322.
[124] Medtronic, GI Genius. https://www.medtronic.com/covidien/en-us/products/gastrointestinal-artificial-intelligence/gi-genius-intelligent-endoscopy.html.
[125] Edakkanambeth Varayil J, Enders F, Tavanapong W, Oh J, Wong J, de Groen PC. Colonoscopy: what endoscopists inspect under optimal conditions. Gastroenterology 2011;140(5):S–718. https://doi.org/10.1016/s0016-5085(11)62982-x.
[126] Mcdermott MBA, Wang S, Marinsek N, Ranganath R, Ghassemi M, Foschini L. Reproducibility in machine learning for health. Seventh international conference on learning representations ICLR2019 2019http://arxiv.org/abs/1907.01463v1.
[127] Sánchez-Peralta LF, Calderón ÁJ, Cabezón V, Ortega-Morán JF, Sánchez-Margallo FM, Polo F, et al. Systematic acquisition and annotation of clinical cases for the generation of a medical image database. Br J Surg 2019;106(S2):16.
[128] Joskowicz L, Cohen D, Caplan N, Sosna J. Inter-observer variability of manual contour delineation of structures in CT. Eur Radiol 2019;29(3):1391–9. https://doi.org/10.1007/s00330-018-5695-5.
[129] Freedman DA. Bootstrapping regression models. Ann Stat 1981;9(6):1218–28. https://doi.org/10.1214/aos/1176345638.
[130] Picon A, Irusta U, Álvarez-Gila A, Aramendi E, Alonso-Atienza F, Figuera C, et al. Mixed convolutional and long short-term memory network for the detection of lethal ventricular arrhythmia. PLoS One 2019;14(5):e0216756. https://doi.org/10.1371/journal.pone.0216756.
[131] He B, Guan Y, Dai R. Classifying medical relations in clinical text via convolutional neural networks. Artif Intell Med 2019;93:43–9. https://doi.org/10.1016/j.artmed.2018.05.001.
[132] Kooi T, Litjens G, Van Ginneken B, Gubern-Mérida A, Sánchez CI, Mann R, et al. Large scale deep learning for computer aided detection of mammographic lesions. Med Image Anal 2017;35:303–12. https://doi.org/10.1016/j.media.2016.07.007.
[133] Shin Y, Qadir HA, Balasingham I. Abnormal colon polyp image synthesis using conditional adversarial networks for improved detection performance. IEEE Access 2018;6:56007–17. https://doi.org/10.1109/ACCESS.2018.2872717.
[134] Ranganathan P, Pramesh CS, Aggarwal R. Common pitfalls in statistical analysis: measures of agreement. Perspect Clin Res 2017;8(4):187–91. https://doi.org/10.4103/picr.PICR.
[135] Robinson R, Valindria VV, Bai W, Oktay O, Kainz B, Suzuki H, et al. Automated quality control in image segmentation: application to the UK Biobank Cardiac MR Imaging study. J Cardiovasc Magn Reson 2019:1–17. https://doi.org/10.1186/s12968-019-0523-x.
[136] Cubuk ED, Zoph B, Mane D, Vasudevan V, Le QV. AutoAugment: learning augmentation policies from data. 2018http://arxiv.org/abs/1805.09501.
[137] Tran T, Pham T, Carneiro G, Palmer L, Reid I. A Bayesian data augmentation approach for learning deep models. 31st conference on neural information processing systems (NIPS 2017) 2017http://arxiv.org/abs/1710.10564.
[138] Bernal J, Sánchez FJ, Rodríguez de Miguel C, Fernández-Esparrach G. Building up the future of colonoscopy? A synergy between clinicians and computer scientists. Screening for colorectal cancer with colonoscopy. InTech; 2015. https://doi.org/10.5772/61012.
[139] Pineau J. The machine learning reproducibility checklist (version 1.0). 2019https://www.cs.mcgill.ca/jpineau/ReproducibilityChecklist.pdf.

8.3 Unravelling the effect of data augmentation transformations in polyp segmentation

Dr. J. Blas Pagador Carrasco y Dr. Artzai Picón Ruiz, como directores de la tesis titulada "Mejoras en la detección, localización y segmentación automática de pólipos colorrectales basadas en técnicas de aprendizaje profundo", certificamos el factor de impacto y la categorización de la siguiente publicación, incluida en la tesis doctoral. Del mismo modo, se especifica la aportación del doctorando.

Autores: **Luisa F. Sánchez-Peralta**, Artzai Picón, Francisco M. Sánchez-Margallo, J. Blas Pagador

Título: Unravelling the effect of data augmentation transformations in polyp segmentation

Revista: International Journal of Computer Assisted Radiology and Surgery

Otros datos: vol 15, págs. 1975-1988, año 2020

DOI: 0.1007/s11548-020-02262-4

Factor de impacto (2020): 2.924 (Q2 – Radiology, Nuclear Medicine & Medical Imaging, Q2 – Surgery, Q3 – Engineering, Biomedical)

Abstract: Purpose Data augmentation is a common technique to overcome the lack of large annotated databases, a usual situation when applying deep learning to medical imaging problems. Nevertheless, there is no consensus on which transformations to apply for a particular field. This work aims at identifying the effect of different transformations on polyp segmentation using deep learning. Methods A set of transformations and ranges have been selected, considering image-based (width and height shift, rotation, shear, zooming, horizontal and vertical flip and elastic deformation), pixel-based (changes in brightness and contrast) and application-based (specular lights and blurry frames) transformations. A model has been trained under the same conditions without data augmentation transformations (baseline) and for each of the transformation and ranges, using CVCEndoSceneStill and Kvasir-SEG, independently. Statistical analysis is performed to compare the baseline performance against results of each range of each transformation on the same test set for each dataset. Results This basic method identifies the most adequate transformations for each dataset. For CVC-EndoSceneStill, changes in brightness and contrast significantly improve the model performance. On the contrary, Kvasir-SEG benefits to a greater extent from the image-based transformations, especially rotation and shear. Augmentation with synthetic specular lights also improves the performance. Conclusion Despite being infrequently used, pixel-based transformations show a great potential to improve polyp segmentation in CVC-EndoSceneStill. On the other hand, image-based transformations are more suitable for Kvasir-SEG. Problem-based transformations behave similarly in both datasets. Polyp area, brightness and contrast of the dataset have an influence on these differences.

Contribución del doctorando: Conceptualization, methodology, software, validation, formal analysis, investigation, data curation, writing—original draft preparation, writing—review and editing, visualization.

Dr. D. José Blas Pagador Carrasco Dr. D. Artzai Picón Ruiz

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

International Journal of Computer Assisted Radiology and Surgery
https://doi.org/10.1007/s11548-020-02262-4

ORIGINAL ARTICLE

Unravelling the effect of data augmentation transformations in polyp segmentation

Luisa F. Sánchez-Peralta[1] · Artzai Picón[2] · Francisco M. Sánchez-Margallo[1] · J. Blas Pagador[1]

Received: 13 January 2020 / Accepted: 14 September 2020
© The Author(s) 2020

Abstract

Purpose Data augmentation is a common technique to overcome the lack of large annotated databases, a usual situation when applying deep learning to medical imaging problems. Nevertheless, there is no consensus on which transformations to apply for a particular field. This work aims at identifying the effect of different transformations on polyp segmentation using deep learning.

Methods A set of transformations and ranges have been selected, considering image-based (width and height shift, rotation, shear, zooming, horizontal and vertical flip and elastic deformation), pixel-based (changes in brightness and contrast) and application-based (specular lights and blurry frames) transformations. A model has been trained under the same conditions without data augmentation transformations (baseline) and for each of the transformation and ranges, using CVC-EndoSceneStill and Kvasir-SEG, independently. Statistical analysis is performed to compare the baseline performance against results of each range of each transformation on the same test set for each dataset.

Results This basic method identifies the most adequate transformations for each dataset. For CVC-EndoSceneStill, changes in brightness and contrast significantly improve the model performance. On the contrary, Kvasir-SEG benefits to a greater extent from the image-based transformations, especially rotation and shear. Augmentation with synthetic specular lights also improves the performance.

Conclusion Despite being infrequently used, pixel-based transformations show a great potential to improve polyp segmentation in CVC-EndoSceneStill. On the other hand, image-based transformations are more suitable for Kvasir-SEG. Problem-based transformations behave similarly in both datasets. Polyp area, brightness and contrast of the dataset have an influence on these differences.

Keywords Polyp segmentation · Deep learning · Data augmentation · Transformations · Semantic segmentation

Introduction

Deep learning techniques have been widely used for the last years as they have proved their ability to extract features for different computer vision tasks such as object detection, classification or segmentation [1]. Undoubtedly, these techniques have also been used for medical imaging with great success [2, 3]. Even though, one limitation that must be faced in this field is the lack of large datasets with relevant annotations and/or labelling [4, 5]. One of the most widely used strategies for addressing this problem is data augmentation [6].

Data augmentation for images consists of increasing the amount and diversity of training cases based on the available images in the database through the application of image transformations such as translation or flipping of the original image [7]. Different computational libraries have been created to perform these transformation functions [8, 9]. However, the selection of the most suitable strategy remains a trial-and-error process that depends on the experience, imagination and time of the researcher [10]. There are several studies analysing the effect of data augmentation for image

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11548-020-02262-4) contains supplementary material, which is available to authorized users.

✉ Luisa F. Sánchez-Peralta
lfsanchez@ccmijesususon.com

1 Jesús Usón Minimally Invasive Surgery Centre, Road N-521, km 41.8, 10071 Cáceres, Spain

2 Tecnalia Research and Innovation, Zamudio, Spain

Published online: 28 September 2020

classification tasks [11–14], but this field is not fully explored for semantic segmentation yet [15].

Computer-assisted diagnosis (CAD) systems for early detection of colorectal cancer have also benefited from the application of deep learning techniques [16–18]. Publicly available datasets range from hundreds of images with a manually segmented binary mask, such as CVC-EndoSceneStill [19] or Kvasir-SEG [20], to thousands of video frames with an approximated elliptical binary mask, such as CVC-VideoClinicDB [21, 22]. For polyp segmentation, it is easy to find several works in which data augmentation has been used. Nevertheless, there is a wide variety of transformations selected as well as their ranges (for example, rotating between − 45° and 45° instead of between − 90° and 90°). Table 1 gathers the applied transformations and their ranges, when available, for recent works on polyp segmentation using deep learning. Although there are authors who do use data augmentation, they do not describe the transformations applied [23]. Besides, it is also important to point out that more intense data augmentation does not necessarily yield to increased performance [24]. The particularities of the medical image type must also be taken into consideration for selecting data augmentation transformations, as the image might have particularities that affect image processing methods. For polyp segmentation, specular lights negatively affect detection methods as they prominently appear, hiding colour and textural information [25].

We hypothesize that the application of different transformations as well as different ranges for the same transformation might lead to differences in performance. Thus, the objective of this work is to elucidate the effect of different image transformations and their ranges used for data augmentation for polyp segmentation. Therefore, this work does not pursue to obtain the best segmentation results but to analyse how the different transformations and their ranges used in data augmentation might influence the results of polyp segmentation in endoscopic images using deep learning.

Table 1 Transformations used for data augmentation in polyp segmentation

Work	Year	Rotation	Width shift	Height shift	Shear	Zoom	Flip	Warp	Gaussian noise	Contrast	Brightness	Patch selection
Jha [20]	2020		–	–	–	✓	✓	–	–	–	✓	–
Guo [26]	2019	✓	–	–	–	✓	✓	–	–	–	✓	–
Kang [27]	2019	(− 45°, 45°)	–	–	(− 16°, 16°)	(0.5, 1.5)	✓	–	–	(0.5, 1.5)	(0.8, 1.5)	✓
Akbari [28]	2018	10° interval, between 0°–290°	–	–	–	–	✓	–	–	–	–	15 patches/image
Brandao [29]	2018	–	–	–	–	–	✓	–	–	–	–	224 × 224 patches
Wichakam [30]	2018	up to 180°	(0, 20%)	(0, 20%)	up to 20%	(−0.8, 1.2)	✓	–	–	–	–	–
Wickstrom [31]	2018	(−90°, 90°)	–	–	(0, 0.4)	(0.8, 1.2)	–	–	–	–	–	224 × 224 patches
Bardhi [32]	2017	✓	✓	✓		–	✓	–	✓	–	–	✓
Li [33]	2017	✓	✓	✓	–	–	–	–	✓	✓	–	–
Vázquez [19]	2017	(0°, 180°)	–	–	(0, 0.4)	(0.9, 1.1)	–	(0, 10)	–	–	–	–

Methods

Transformations

Different transformations have been considered in this study, which can be classified into three categories. For each transformation, a suitable range of values has been established (Table 2). Figure 1 shows an example of the result of applying each transformation to an image. In the case of image-based transformations, image and mask are transformed in the same way.

To model the specular lights, the CVC-EndoSceneStill database [19] has been used, as it provides a manually segmented class for specular lights in endoscopic images.

Table 2 Transformations and ranges analysed in this study

Transformation	Parameter definition	Ranges	Total cases
Image-based transformations			
Width shift	% of the image displaced to the right or to the left	0–90%, with 10% intervals	9 cases
Height shift	% of the image displaced up or down	0–90%, with 10% intervals	9 cases
Rotation	± Degrees that the image is rotated	0–180°, with up to 45° intervals	8 cases
Shear	± Shear angle in counter-clockwise direction	0–180°, with up to 45° intervals	8 cases
Zoom out	Factor by which the image size is multiplied	$1 - x$, $x \in [0.1, 0.9]$, with 0.1 intervals	9 cases
Zoom in	Factor by which the image size is multiplied	$1 + x$, $x \in [0.1, 1.0]$, with 0.1 intervals	10 cases
Flip	Vertically and horizontally flip the image	True	2 cases
Elastic deformation	Parameters as indicated in [32]	α values: 250, 500, 1000, 2000, 3000, 4000, 5000, 6000 σ value: fixed at 40	8 cases
Pixel-based transformations			
Brightness	± value to be added to the actual pixel value for all RGB channels equally	[25, 175], with 25 intervals	7 cases
Brightness	Value to be added to the actual pixel value for each RGB channel independently	[25, 175], with 25 intervals	7 cases
Contrast	Value to multiply the actual pixel value for all RGB channels equally	$[1 - x, 1 + \mathrm{x}x, x \in [0.2, 1.0]$, in intervals of 0.2	5 cases
Contrast	Value to multiply the actual pixel value for each RGB channel independently	$[1 - x, 1 + x], x \in [0.2, 0.8]$, in intervals of 0.2	4 cases
Application-based transformations			
Specular lights	Overexposed light ellipses simulating the effect of bright points	True	1 case
Blurry images	Window size of a mean filter	[1, 15], only even integers	7 cases

Specular lights are modelled as ellipses of variable size and orientation. Size of major and minor axes are obtained from the specular lights in CVC-EndoSceneStill database, corresponding to a mean major axis of 7.77 ± 10.36 pixels (range 0–259.81) and a mean minor axis of 3.82 ± 4.29 pixels (range 0–137.39). The number of specular lights per image is modelled as a positive left-skewed distribution, with mean 18.20 and standard deviation 16.97, according to the distribution of CVC-EndoSceneStill. In the image, pixel values are set to 255 in all channels to create the ellipses according to the previously described distributions, with random locations on the image.

Datasets, architecture and training process

Two publicly available datasets have been used in this work. CVC-EndoSceneStill [19] contains 912 images obtained from 44 video sequences collected from 36 patients. It explicitly indicates the images belonging to the training, validations and test sets. In this work, this division has been used. This way, all experiments use the same images, which allows for a fair comparison of performance. The training, validation and test sets comprise 547, 183 and 182 images, respectively. The second dataset is Kvasir-SEG [20]. It provides 1000 polyp images. The dataset has been divided into training, valida-

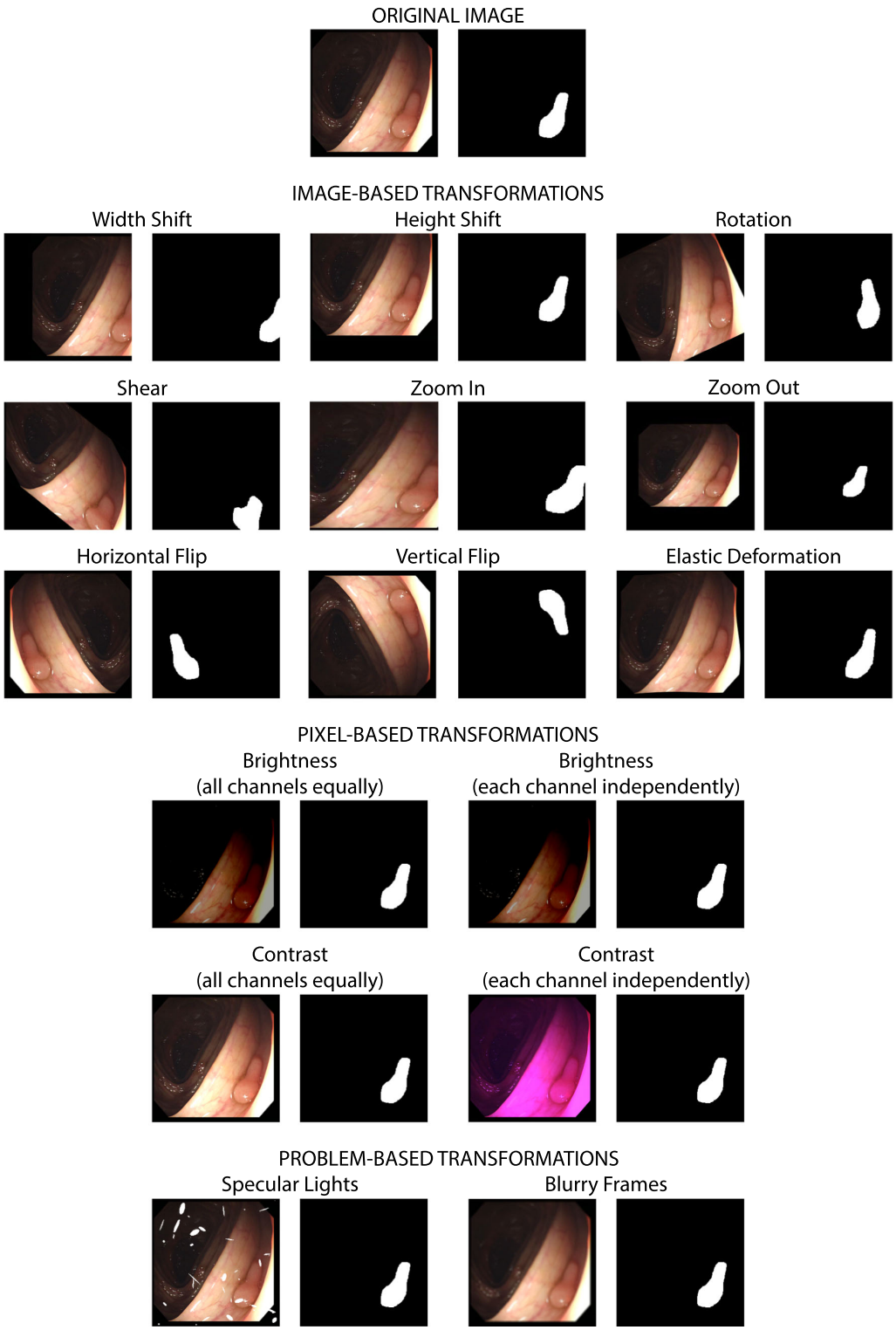


Fig. 1 Original and transformed images

Table 3 Details for the datasets used in this study

	CVC-EndoSceneStill	Kvasir-SEG
Void area (%)	23.73 ± 5.57 (27.83–14.62)	15.23 ± 4.82 (28.44–6.16)
Polyp area relative to the valid area (%)	12.50 ± 11.49 (66.15–0.75)	17.36 ± 15.65 (83.66–0.61)
Mean value of brightness channel in HSV [34]	0.560 ± 0.006 (1.000–0.000)	0.622 ± 0.003 (1.000–0.000)
Histogram flatness measure [35]	0.858 ± 0.121 (0.959–0.000)	0.419 ± 0.443 (0.962–0.000)
Histogram spread [35]	0.252 ± 0.088 (0.520–0.076)	0.218 ± 0.070 (0.432–0.075)

Results are reported as mean ± standard deviation. Minimum and maximum values are indicated between brackets. The void area refers to the black area in the images, while the remaining area is considered as valid area

tion and test sets (800, 200 and 200 images, respectively), as this division is not provided by the dataset's owners. Both datasets provide binary masks for each polyp image, where pixels corresponding to the class are labelled with 1, and 0 otherwise. Each dataset is used on its own to replicate the same experiments for further comparison of results. Table 3 shows some characteristics of the images included in the test sets of the datasets. Kvasir-SEG presents bigger polyps than CVC-EndoSceneStill, with images that are brighter and with more contrast and where the void area is smaller.

Our network architecture (Fig. 2) is based on a U-Net architecture [36]. The down-sampling path transforms the input image of size $256 \times 256 \times 3$ to a feature map of $16 \times 16 \times 1024$ by applying five convolutional blocks. These blocks consist of two 3×3 convolutional layers, each one with a rectified linear unit, and a 2×2 max pool layer, except for the last block. The up-sampling path includes four blocks that produce a $256 \times 256 \times 1$ probability map. Each block starts with a 2×2 up-sampling layer followed by a 3×3 convolutional layer, to whose result the corresponding feature map from the down-sampling path is concatenated. Zero padding preserves sizes along convolutional layers. We included batch normalization both in down- and up-sampling paths.

The network has been implemented using Keras [37] and Tensorflow [38] as backend. Experiments were run on a NVIDIA GTX 1080 GPU with 8 GB memory. The network has been pretrained using CVC-VideoClinicDB [21, 22], whose polyp masks are not precise but approximated to elliptical shapes. The datasets in "Transformations" section are then used to finetune this pretrained model with fixed parameters for all experiments:

- Adam optimizer, with default parameters in Keras: amsgrad = false; beta_1 = 0.9 and beta_2 = 0.999
- Learning rate: starting at 10^{-4}, decreasing to half each epoch and recovering to 10^{-4} each 5 epochs
- 15 epochs
- Batch size: 4
- Image input size: $256 \times 256 \times 3$
- Dropout: 0.5

Each experiment has been repeated ten times to minimize the effect of randomly applying transformations. Results are shown in terms of mean ± standard deviation of the mean. A baseline level has been established by finetuning the model without applying any data augmentation.

Since semantic segmentation is performed through a pixel-wise classification, we face an unbalanced dataset where the negative class (no polyp) is more present than the positive one (polyp) in each image. Therefore, the selected loss function combines the binary cross-entropy and the Jaccard index as in [39]:

$$\text{Loss} = -\frac{1}{n}\sum_{i,j}\left(y_{i,j}\log\hat{y}_{i,j} + \left(1+\hat{y}_{i,j}\right)\right)\log\left(1-\hat{y}_{i,j}\right) - \log J,$$

where the first term corresponds to the binary cross-entropy, being $y_{i,j}$ the ground truth class for pixel (i, j) and $\hat{y}_{i,j}$ the predicted class; and J is the Jaccard index or Intersection over Union (IoU) defined as a similarity measure between sets A and B as:

$$J = IoU(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|},$$

where $|X| = \sum_i x_i$ being x_i is the i-th element of set X; $\cap$ is the intersection of sets and $\cup$ is the union of sets.

Statistical analysis

Results of the ten repetitions have been statistically analysed to identify differences between distributions, using R (version 3.6.1) and RStudio (version 1.2.5033). Permutation test [40] is selected as no assumption on the distributions is required. In the permutation test, firstly the "observed mean" is calculated as the difference between means for the baseline and the group under study. Data are then shuffled and randomly assigned to each group and the corresponding "calculated mean" is obtained as the difference between means of the two groups. After 10000 repetitions, the p value is determined as the percentage of calculated means which are greater than the observed mean. Significance is evaluated at p

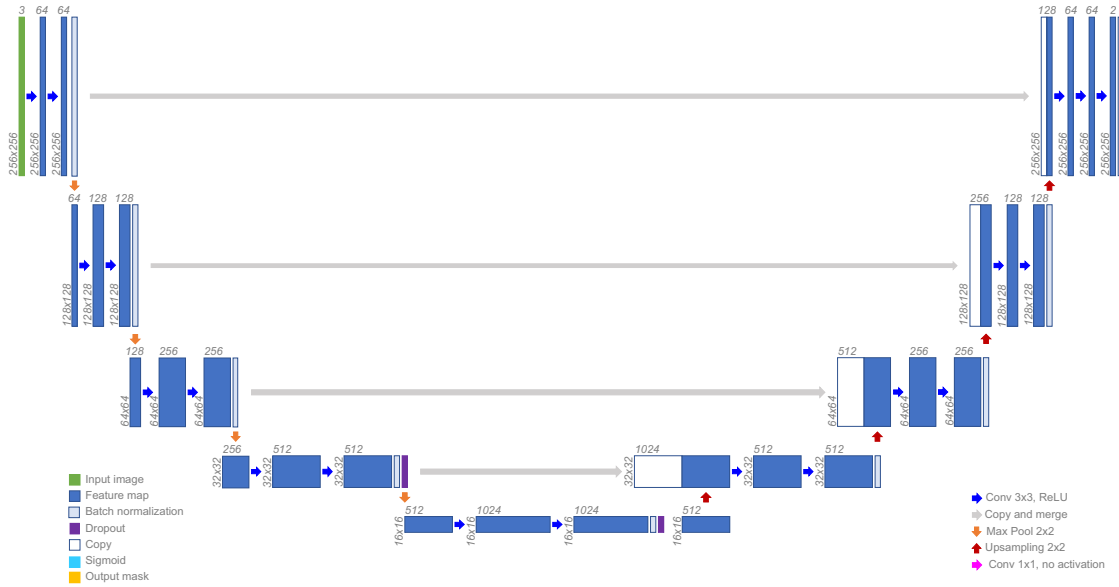


Fig. 2 Network architecture. Figure based on [36]

value < 0.05, p value < 0.01 and p value < 0.001. This analysis is performed for each dataset independently.

Results

For both datasets, Table 4 shows the results for the baseline and all transformations and ranges, together with the results of the permutation test to establish statistically significant differences between baseline and transformations.

Figures 3, 4 and 5 show the range with the highest mean for each transformation for the CVC-EndosceneStill and Kvasir-SEG. Figures for all transformations and ranges can be found in the Supplementary material 1 for CVC-EndoSceneStill dataset and Supplementary material 2 for Kvasir-SEG. All figures show boxplots combined with violin plots, representing the distribution of the results. In these violin plots, the ideal outcome is that the distribution presents a peak at 1. Therefore, the more the distribution looks alike this peak, the better the performance is.

Image-based transformations have different behaviours depending on the dataset, transformation and range. In first place, width and height shift transformations are dependent on the range to either improve or hinder performance of the network in both cases. Only ranges over 40% produce a positive effect, up to 6.59 points, although statistical significance is not achieved in CVC-EndoSceneStill. If Kvasir-SEG is considered, these transformations improve the baseline if small ranges are used, but not significantly. Secondly, rotation and shear results are in all cases under the baseline threshold, reaching 4.43 points decrement in the performance for CVC-EndoSceneStill. On the contrary, these transforms improve performance on Kvasir-SEG in up to 3.41 points, being the greatest improvement in this dataset. Zooming the image has different results depending on whether it is zoom in or out in CVC-EndoSceneStill. Zooming in decreases the performance more than 3.5 points, while zooming out can improve results up to almost 5 points, but significance is not achieved. On the contrary, some ranges from both transforms improve performance in Kvasir-SEG although not significantly. In relation to flipping the image, when CVC-EndoSceneStill is considered, horizontally flipping the image hinders the performance but if flipping is vertical, then performance is increased. In both cases, changes are not significant. On the contrary, both transforms improve performance in Kvasir-SEG, also without statistical significance. Lastly, elastic deformation of the image leads to deterioration of performance of up to 4.45 points in CVC-EndoSceneStill, but improve performance in 1.72 points in Kvasir-SEG.

The second group of transformations modified the pixel-value. On one hand, changes in brightness in CVC-EndoSceneStill, regardless of modifying all channels equally or each channel independently, yield to a better performance of the model of more than 12 points, obtaining significant

Table 4 Mean and standard deviation of the mean for transformations and ranges analysed in both datasets

Transformation	Range	IoU on test set CVC-EndoSceneStill	IoU on test set Kvasir-SEG
None	N/A	59.10 ± 9.35	66.45 ± 8.08
Image-based transformations			
Width Shift	± 10%	60.78 ± 8.99	67.09 ± 7.96
	± 20%	59.45 ± 9.80	**67.34 ± 8.06**
	± 30%	59.31 ± 9.08	66.28 ± 8.22
	± 40%	62.70 ± 8.57	65.94 ± 8.22
	± 50%	62.80 ± 8.84	66.23 ± 8.09
	± 60%	63.02 ± 8.78	66.90 ± 7.86
	± 70%	63.03 ± 8.67	66.82 ± 7.87
	± 80%	61.34 ± 8.62	65.41 ± 7.92
	± 90%	**65.68 ± 8.12***	65.82 ± 7.72
Height shift	± 10%	58.82 ± 8.97	67.00 ± 7.98
	± 20%	58.94 ± 8.80	67.12 ± 8.08
	± 30%	61.81 ± 8.74	**67.26 ± 7.87**
	± 40%	**62.03 ± 8.57**	67.23 ± 7.80
	± 50%	61.78 ± 8.42	67.17 ± 7.89
	± 60%	60.21 ± 8.64	66.97 ± 7.94
	± 70%	61.55 ± 8.46	66.69 ± 7.98
	± 80%	60.42 ± 8.19	66.26 ± 7.94
	± 90%	61.52 ± 8.27	67.06 ± 7.58
Rotation	± 3°	57.74 ± 9.37	66.41 ± 8.09
	± 6°	**59.97 ± 9.06**	65.61 ± 8.16
	± 10°	55.40 ± 9.75	65.74 ± 8.15
	± 15°	55.50 ± 9.65	67.03 ± 8.10
	± 45°	54.66 ± 9.62	68.38 ± 8.00
	± 90°	57.62 ± 9.37	**69.86 ± 7.79**
	± 135°	58.60 ± 9.49	68.22 ± 8.07
	± 180°	58.19 ± 9.35	68.78 ± 8.10
Shear	± 3°	59.62 ± 9.05	66.24 ± 8.11
	± 6°	**61.66 ± 8.98**	67.00 ± 8.02
	± 10°	59.42 ± 9.00	67.32 ± 7.90
	± 15°	57.91 ± 9.10	67.11 ± 7.97
	± 45°	59.07 ± 9.80	**68.88 ± 7.74**
	± 90°	56.38 ± 9.3	67.84 ± 7.85
	± 135°	55.22 ± 9.37	67.53 ± 7.91
	± 180°	57.09 ± 8.89	67.67 ± 7.90
Zoom in	0.9, 1	**60.19 ± 8.54**	66.71 ± 8.08
	0.8, 1	59.98 ± 8.53	67.45 ± 8.01
	0.7, 1	57.01 ± 9.46	67.56 ± 8.24
	0.6, 1	55.57 ± 10.07	68.54 ± 8.14
	0.5, 1	57.37 ± 10.30	**68.80 ± 8.25**
	0.4, 1	58.58 ± 10.18	67.26 ± 8.29
	0.3, 1	58.41 ± 10.40	66.54 ± 8.30
	0.2, 1	57.71 ± 10.34	65.54 ± 8.37
	0.1, 1	57.56 ± 10.06	64.05 ± 8.51
Zoom out	1, 1.1	58.70 ± 9.12	65.48 ± 8.17
	1, 1.2	61.64 ± 8.26	66.25 ± 8.09
	1, 1.3	58.99 ± 8.50	65.88 ± 8.03
	1, 1.4	62.21 ± 8.04	66.13 ± 7.98
	1, 1.5	61.83 ± 8.39	66.56 ± 7.86
	1, 1.6	**64.03 ± 8.26**	67.38 ± 7.80
	1, 1.7	60.67 ± 7.90	67.38 ± 7.83
	1, 1.8	62.01 ± 8.20	**67.97 ± 7.63**
	1, 1.9	62.73 ± 8.00	67.91 ± 7.57
	1, 2.0	64.00 ± 8.13	67.97 ± 7.64
Horizontal flip	True	55.89 ± 9.22	67.57 ± 8.11
Vertical flip	True	59.54 ± 8.90	67.23 ± 8.08
Elastic deformation	250, 40	**60.26 ± 8.79**	65.92 ± 8.19
	500, 40	59.17 ± 9.31	65.86 ± 8.18
	1000, 40	57.93 ± 9.12	66.97 ± 8.00
	2000, 40	57.83 ± 8.86	67.88 ± 8.02
	3000, 40	55.89 ± 9.14	**68.17 ± 8.00**
	4000, 40	54.65 ± 9.12	66.96 ± 8.20
	5000, 40	56.55 ± 9.13	65.17 ± 8.36
	6000, 40	55.90 ± 9.37	65.02 ± 8.28
Pixel-based transformations			
Brightness, all channels equally	± 25	59.89 ± 84	66.87 ± 7.66
	± 50	63.27 ± 8.41	66.22 ± 7.74
	± 75	66.79 ± 8.28**	65.17 ± 7.76
	± 100	67.99 ± 8.23**	64.55 ± 7.86
	± 125	68.98 ± 7.90***	63.95 ± 7.87
	± 150	**70.07 ± 7.75*****	67.25 ± 7.86
	± 175	68.32 ± 7.74**	**67.70 ± 7.88**
Brightness, each channel independently	± 25	**71.21 ± 7.69*****	67.85 ± 7.84
	± 50	70.90 ± 7.81***	68.28 ± 7.78
	± 75	69.26 ± 8.19***	68.91 ± 7.60
	± 100	69.07 ± 8.26***	69.21 ± 7.51
	± 125	67.86 ± 8.27**	**69.36 ± 7.46**
	± 150	67.86 ± 7.77**	67.07 ± 8.05
	± 175	66.15 ± 8.16*	68.39 ± 7.65

Table 4 continued

Transformation	Range	IoU on test set CVC-EndoSceneStill	IoU on test set Kvasir-SEG
Contrast, all channels equally	0.8, 1.2	58.11 ± 9.35	66.89 ± 7.98
	0.6, 1.4	61.55 ± 8.76	67.31 ± 7.85
	0.4, 1.6	66.17 ± 8.37*	67.92 ± 7.56
	0.2, 1.8	**68.38 ± 8.06****	**68.16 ± 7.63**
	0.0, 2.0	60.54 ± 9.43	66.29 ± 8.14
Contrast, each channel independently	0.8, 1.2	71.80 ± 7.61***	**67.68 ± 7.95**
	0.6, 1.4	71.70 ± 7.62***	66.58 ± 7.79
	0.4, 1.6	**72.34 ± 7.81*****	66.45 ± 7.63
	0.2, 1.8	70.54 ± 7.97***	66.83 ± 7.46
Application-based transformations			
Specular lights	True	59.64 ± 9.06	67.52 ± 7.59
Blurry image	3	**60.32 ± 8.67**	66.14 ± 8.01
	5	58.94 ± 9.37	65.54 ± 8.01
	7	53.61 ± 9.33	64.86 ± 8.05
	9	50.39 ± 9.84**	**66.81 ± 7.93**
	11	51.24 ± 10.02*	64.78 ± 8.12
	13	52.21 ± 9.75*	65.85 ± 8.13
	15	48.41 ± 10.32**	64.91 ± 8.13

Best value for each transformation is indicated in bold
Statistical differences between baseline and the particular case are identified with permutation test
***p value < 0.001; **p value < 0.01; *p value < 0.05

differences in all cases but two. Similarly, modifying the contrast reached an increment of 13.25 points with respect to the baseline, being this the greatest improvement in all transformations and ranges, and obtaining statistically significant differences for all ranges if channels are modified independently and two out of four if they are equally modified. This behaviour is not so strong in the Kvasir-SEG, while changing brightness and contrast do improve performance in some ranges, significance is not achieved.

Lastly, we analysed transformations based on specific problems of colonoscopy images: adding specular lights and blurring frames. In the first case, including specular lights increased performance in half point and one point regarding the baseline for each dataset, although significance is not achieved in any dataset. On the second case, blurring the image resulted on a significant decrement of up to 10.69 points when compared to the baseline in the case of CVC-EndoScenestill, but only 1.59 points and no significance in Kvasir-SEG.

Based on these results, we have also analysed combinations of transformation for the different datasets. Results are included in Table 5 and Fig. 6. In all cases for CVC-EndoSceneStill, the mean of these combinations is similar to the transformation with higher mean, but the distributions are improved as the 25 quartile is increased and the standard deviation is minimized. On the other hand, the combination of all image-based transformations hinders the performance, proving that more data augmentation is not always better [24], as only the two image-based transformations with higher mean obtain the best results.

Discussion and conclusion

Data augmentation is a useful tool to increase the number of training samples when the available dataset is scarce, a situation that is well-known when using medical images. The effect of different transformations usually applied in data augmentation for polyp segmentation has yet to be rigorously analysed. In this work, we have found that although image-based transformations are usually applied in the state of the art, pixel-based transformations produce better results for CVC-EndoSceneStill. These transformations modify the particular value of the pixel, so the model is invariant to colour information, which improves its generalization capacity. On the other hand, Kvasir-SEG benefits to a greater extent from the image-based transformations.

In the light of the results, four new groups of transformations can be established:

1. Transformations that always improve the performance in CVC-EndoScenStill and Kvasir-SEG: vertical flip, changes on brightness for each channel independently, changes on contrast (all channels equally and each channel independently) and inclusion of specular lights. All these transformations improve the performance over the baseline, although statistical significance is mainly found in changes of brightness and contrast in CVC-EndoSceneStill.
2. Transformations that always hinder the performance in CVC-EndoScenStill and Kvasir-SEG: elastic deformation and blurry frames (mean filter). While blurry frames could be expected to minimize the performance as they reduce the details in the image, elastic deformation might have been expected to improve performance. Although blurry frames are a common situation during a live

International Journal of Computer Assisted Radiology and Surgery

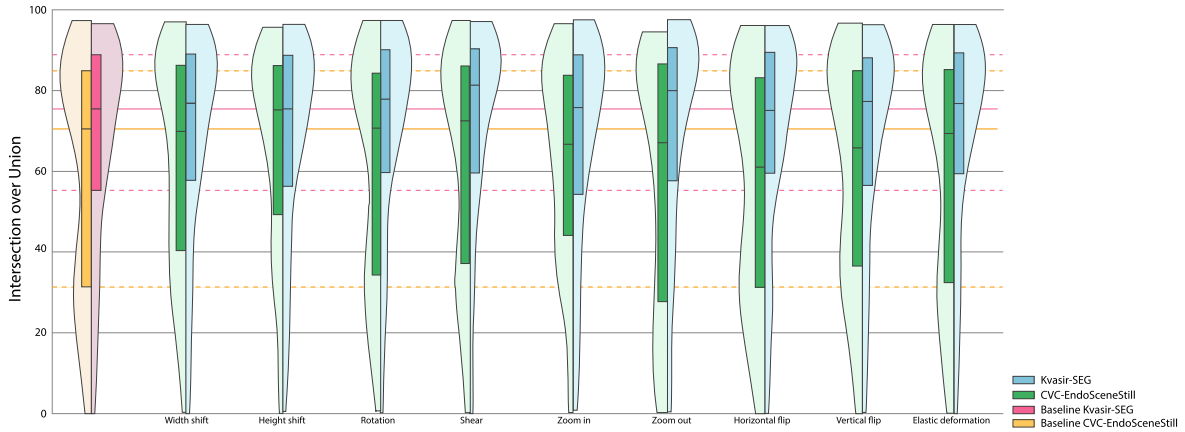


Fig. 3 Results for image-based transformations. Ranges with highest mean are shown for each transformation and dataset. Baselines of each dataset are included. Their median and quartiles are prolongated on the background for reference. For the CVC-EndoSceneStill: ± 90% width shift; ± 40% height shift; ± 6° rotation, ± 45° shear; 0.9 zoom in; 0.4 zoom out; (250,40) elastic deformation. For the Kvasir-SEG: ± 20% width shift; ± 30% height shift; ± 90° rotation, ± 45° shear; 0.5 zoom in; 0.2 zoom out; (3000,40) elastic deformation

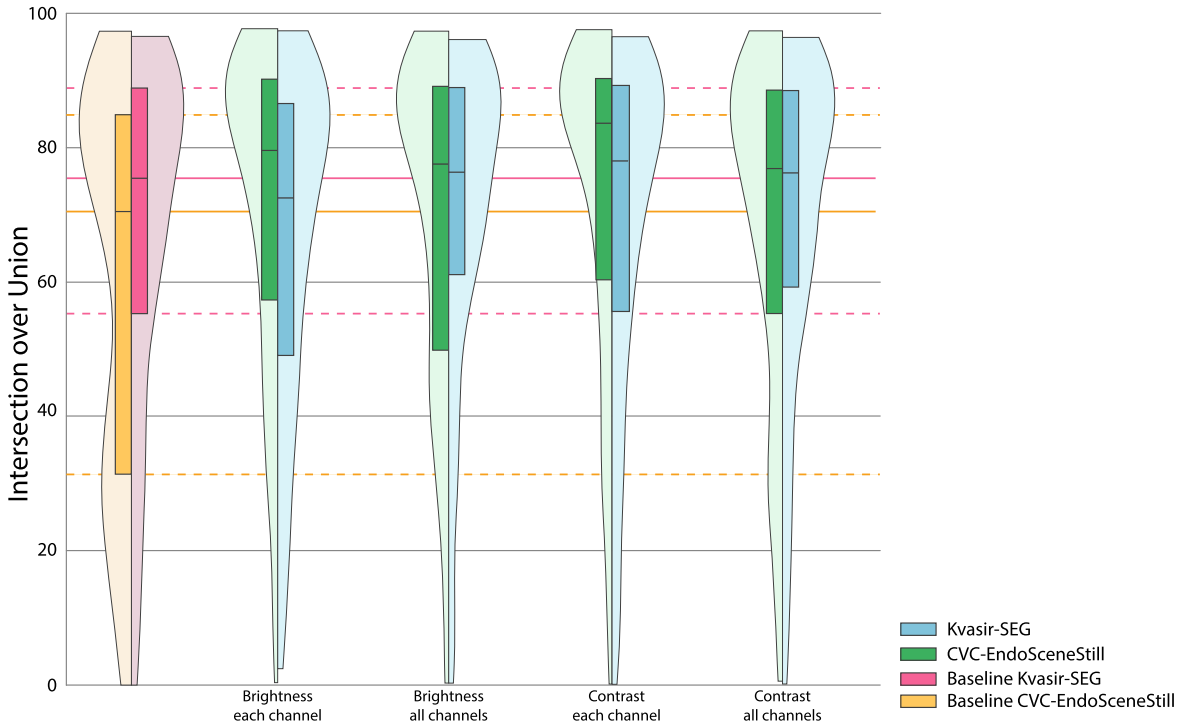


Fig. 4 Results for image-based transformations. Ranges with highest mean are shown for each transformation and dataset. Baselines of each dataset are included. Their median and quartiles are prolongated on the background for reference. For the CVC-EndoSceneStill: ± 150 for brightness in all channels equally; ± 25 for brightness in each channel independently; (0.2–1.8) for contrast in all channels equally; and (0.4–1.6) for brightness in each channel independently. For the Kvasir-SEG: ± 175 for brightness in all channels equally; ± 125 for brightness in each channel independently; (0.2–1.8) for contrast in all channels equally; and (0.8–1.2) for brightness in each channel independently

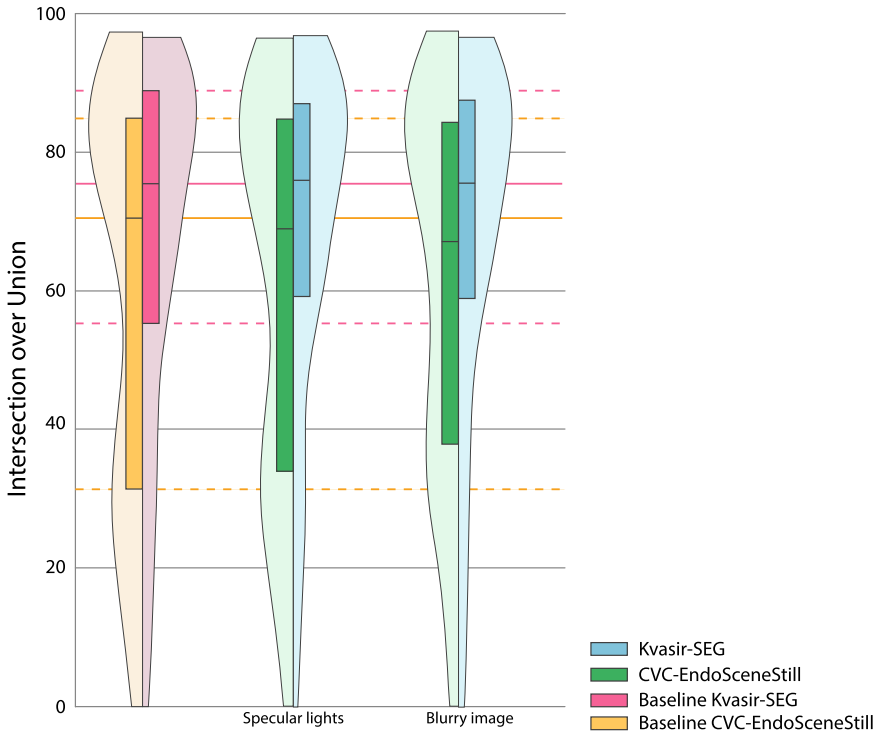


Fig. 5 Results for problem-based transformations. Ranges with highest mean are shown for each transformation and dataset. Baselines of each dataset are included. Their median and quartiles are prolongated on the background for reference. For the CVC-EndoSceneStill: 3 for blurry images. For the Kvasir-SEG: 9 for blurry images

colonoscopy, the inclusion of mean filter as transformation for data augmentation does not improve the final performance of the model. This is probably explained by the use of databases, where frames are previously selected and not blurry frames are included.

3. Transformations whose effect on performance depends on the selected range in CVC-EndoScenStill and Kvasir-SEG: height and width shifts, as well as zoom in and out. In the first two cases, ranges over 40% do contribute to improve performance, while under the threshold either the transformation does not add improvement or decrement the performance. On the other hand, zoom behaviour also depends on the range. Smaller ranges of zoom in and larger ranges of zoom out improve the performance over the baseline, although not always significantly. One reason for the performance of the zoom in might be grounded on the low quality of the original images, resulting in blurry zoomed images. Therefore, when using them for data augmentation, it is recommended to carefully check whether the range is suitable or not.
4. Transformations whose effect on performance depends on the dataset, CVC-EndoScenStill or Kvasir-SEG: This relates mainly to rotation, shear and changes on brightness for all channels equally, and, to a lesser extent, horizontal flip. This might be due to differences in polyp size, void area, brightness and contrast in the images of the two datasets.

In summary, CVC-EndoSceneStill is more prone to benefits of data augmentation if pixel-based transformations are used, as the histogram is flatter, and images are darker than in Kvasir-SEG. On the contrary, image-based transformations appear to be more suitable in Kvasir-SEG, where the void area is smaller, and the polyp occupy a greater area of the valid image. Lastly, problem-based transformations behave similarly in both datasets, as they are rooted on the endoscopic image acquisition. It is also important to mention that the baseline of Kvasir-Seg showed already a better performance than CVC-EndosSceneStill, giving less room for improvement to data augmentation.

There are different approaches to overcome the scarce labelled datasets in medical imaging. On the one side, and in order to increase the size of the training set, a first approach would be to increase the number of annotated samples by experts. In this regard, efforts are been focused on developing tools which facilitates the manual annotation of images, such as GTCreatorTool [22], which is a flexible annotation

Table 5 Mean and standard deviation of combinations analysed

	CVC-EndoSceneStill		Kvasir-SEG	
	Transformations	IoU on test set	Transformations	IoU on test set
Baseline	None	59.10 ± 9.35	None	66.45 ± 8.08
Transformation and range with highest mean for each one of the three types of transforms	Width at ± 90%	72.30 ± 7.26***	90° rotation	65.53 ± 7.98
	Change of contrast: each channel independently, with range [0.4, 1.6]		Change of brightness: each channel independently, with range ± 125	
	Inclusion of specular lights		Inclusion of specular lights	
Range with highest mean of the image-based transformations, provided that they improve the baseline result	Width at ± 90%	65.19 ± 7.81*	Width at ± 20%	57.97 ± 9.21**
	Height at ± 40%		Height at ± 30%	
	Zoom with range [1, 1.6]		90° rotation	
	Vertical flip		45° shear	
			Zoom with range [0.5, 1]	
			Vertical flip	
			Horizontal flip	
			Elastic deformation, with values (3000,40)	
The two transformations with higher mean	Change of contrast: each channel independently, with range [0.4, 1.6]	70.50 ± 7.69***	90° rotation	69.24 ± 7.85
	Change of brightness: each channel independently, with range ± 25		45° shear	

Statistical differences between baseline and combination are identified with permutation test
***p value < 0.001; **p value < 0.01; *p value < 0.05

tool which minimizes annotation time and allows for sharing annotations among experts. Beyond the transformations analysed in this paper, other alternatives would be to add polyps in nonpolypoid samples [41] or more advances approaches such as emulating data augmentation during learning by the image generation through a hetero-encoder [42]. On the other hand, it would be possible to explore alternatives to supervised training, which already seems to provide good results with self-supervised learning [43] or similarity-based active Learning [44].

There are limitations in this study that must be acknowledged. Ideally, it would be necessary to independently analyse all combinations. Since that would mean almost 6 million experiments, alternatives such as AutoAugment [7] or Smart Augmentation [10] would be more suitable for identification of the best combination of transformations. Another possibility could be the application of Bayesian methods [45] or coordinate ascent optimization [46, 47] taking the optimal setting of each transform to identify the best combination. Future work should place emphasis on applying any of these alternatives to the particular field of polyp segmentation. Another limitation is the fact that the experiments have not pursued the best model, so training has been stopped at 15 epochs. It might be possible that with a more extensive training some of the transformations could have showed better results. Nevertheless, 15 epochs is enough training to establish the tendency of the model performance when finetuning it with a small database.

Further research is also possible in this line of work. Future works might focus on the effect of data augmentation on other segmentation approaches, such as the fuzzy C-mean clustering, which has shown good preliminary results on the Kvasir-SEG database [20].

In conclusion, this study shows that different transformations and ranges lead to differences in model performance. Despite not being so frequent as the other types, pixel-based transformations show a great potential to improve polyp segmentation. Augmenting colour variability when training the model allows for a better generalization of the model resulting in better prediction. On the other hand, image-based transformations and their ranges should be carefully selected to not hinder the model performance and obtain the expected benefits of data augmentation.

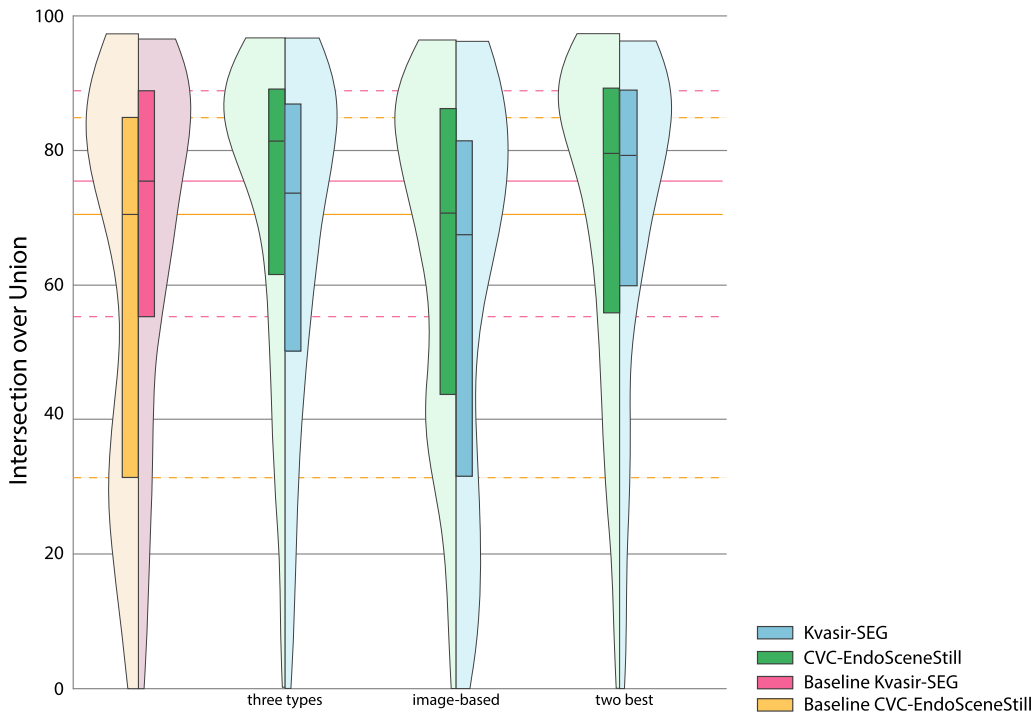


Fig. 6 Results for combination of transformations. Baselines of each dataset are included. Their median and quartiles are prolongated on the background for reference. Combination of the transformation and range with highest mean for each one of the three types of transforms for each dataset. For CVC-EndoSceneStill: width at ± 90%, change of contrast: each channel independently, with range [0.4, 1.6], and inclusion of specular lights. For Kvasir-SEG: 90° rotation, change of brightness: each channel independently, with range ± 125, and inclusion of specular lights. Combination of the range with highest mean of the image-based transformations, provided that they improve the baseline result. For CVC-EndoSceneStill: width at ± 90%, height at ± 40%, zoom with range [1, 1.6], and vertical flip. For Kvasir-SEG: width at ± 20%, height at ± 30%, 90° rotation, 45° shear, zoom with range [0.5, 1], vertical flip, horizontal flip, and elastic deformation, with values (3000,40). Combination of the two transformations with higher mean. For CVC-EndoSceneStill: change of contrast: each channel independently, with range [0.4, 1.6] and change of brightness: each channel independently, with range ± 25. For Kvasir-SEG: 90° rotation and 45° shear

Acknowledgements The authors would like to thank Agueda Azpeitia and Juan Carlos Gómez-Blanco for their statistically significant help.

Funding This work was partially supported by PICCOLO project. This project has received funding from the European Union's Horizon2020 research and innovation programme under Grant Agreement No 732111. The sole responsibility of this publication lies with the author. The European Union is not responsible for any use that may be made of the information contained therein.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

Informed consent This article does not contain patient data.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521:436–444. https://doi.org/10.1038/nature14539
2. Biswas M, Kuppili V, Saba L, Edla DR, Suri HS, Cuadrado-Godia E, Laird JR, Marinhoe RT, Sanches JM, Nicolaides A, Suri JS

(2019) State-of-the-art review on deep learning in medical imaging. Front Biosci 24:380–406. https://doi.org/10.2741/4725
3. Ker J, Wang L, Rao J, Lim T (2018) Deep learning applications in medical image analysis. IEEE Access 6:9375–9389. https://doi.org/10.1109/ACCESS.2017.2788044
4. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, van der Laak JAWM, van Ginneken B, Sánchez CI (2017) A survey on deep learning in medical image analysis. Med Image Anal 42:60–88. https://doi.org/10.1016/j.media.2017.07.005
5. Medela A, Picon A, Saratxaga CL, Belar O, Cabezón V, Cicchi R, Bilbao R, Glover B (2019) Few shot learning in histopathological images: reducing the need of labeled data on biological datasets. In: 2019 IEEE 16th international symposium on biomedical imaging (ISBI). https://doi.org/https://doi.org/10.1109/ISBI.2019.8759182
6. Shorten C, Khoshgoftaar TM (2019) A survey on image data augmentation for deep learning. J Big Data. https://doi.org/10.1186/s40537-019-0197-0
7. Cubuk ED, Zoph B, Mane D, Vasudevan V, Le QV (2018) AutoAugment: learning augmentation policies from data. arXiv. https://arxiv.org/abs/1805.09501. Accessed 5 Jan 2020
8. Jung AB, Wada K, Crall J, Tanaka S, Graving J,Yadav S, Banerjee J, Vecsei G, Kraft A, Borovec J, Vallentin C, Zhydenko S, Pfeiffer K, Cook B, Fernández I, Chi-Hung W, Ayala-Acevedo A, Meudec R, Laporte M (2019) Imgaug. https://github.com/aleju/imgaug. Accessed 5 Jan 2020
9. Casado-García Á, Domínguez C, García-Domínguez M, Heras J, Mata E, Pascual V (2019) CLoDSA: a tool for augmentation in classification, localization, detection, semantic segmentation and instance segmentation tasks. BMC Bioinform 20:1–14. https://doi.org/10.1186/s12859-019-2931-1
10. Lemley J, Bazrafkan S, Corcoran P (2017) Smart augmentation learning an optimal data augmentation strategy. IEEE Access 5:5858–5869. https://doi.org/10.1109/ACCESS.2017.2696121
11. Okafor E, Schomaker L, Wiering MA (2018) An analysis of rotation matrix and colour constancy data augmentation in classifying images of animals augmentation in classifying images of animals. J Inf Telecommun. https://doi.org/10.1080/24751839.2018.1479932
12. Shijie J, Ping W, Peiyi J, Siping H (2017) Research on data augmentation for image classification based on convolution neural networks. In: 2017 Chinese automation congress (CAC). pp 4165–4170
13. Wang J, Perez L (2017) The effectiveness of data augmentation in image classification using deep learning. https://arxiv.org/abs/1712.04621
14. Zhang C, Tavanapong W, Wong J, de Groen PC, Oh JH (2017) Real data augmentation for medical image classification. In: Intravascular imaging and computer assisted stenting, and large-scale annotation of biomedical data and expert label synthesis. pp 67–76. https://doi.org/https://doi.org/10.1007/978-3-319-67534-3_8
15. Ma R, Tao P, Tang H (2019) Optimizing data augmentation for semantic segmentation on small-scale dataset. In: Proceedings of the 2nd international conference on control and computer vision—ICCCV 2019. pp 77–81
16. Urban G, Tripathi P, Alkayali T, Mittal M, Jalali F, Karnes W, Baldi P (2018) Deep learning localizes and identifies polyps in real time with 96% accuracy in screening colonoscopy. Gastroenterology 155:1069–1078.e8. https://doi.org/10.1053/j.gastro.2018.06.037
17. Wang P, Berzin TM, Glissen Brown JR, Bharadwaj S, Becq A, Xiao X, Liu P, Li L, Song Y, Zhang D, Li Y, Xu G, Tu M, Liu X (2019) Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: a prospective randomised controlled study. Gut. https://doi.org/10.1136/gutjnl-2018-317500
18. Alagappan M, Brown JRG, Mori Y, Berzin TM (2018) Artificial intelligence in gastrointestinal endoscopy: the future is almost here. World J Gastrointest Endosc 10:239–249. https://doi.org/10.4253/wjge.v10.i10.239
19. Vázquez D, Bernal J, Sánchez FJ, Fernández-Esparrach G, López AM, Romero A, Drozdzal M, Courville A (2017) A benchmark for endoluminal scene segmentation of colonoscopy images. J Healthc Eng. https://doi.org/10.1155/2017/4037190
20. Jha D, Smedsrud PH, Riegler MA, Halvorsen P, Johansen HD, de Lange T, Johansen D (2020) Kvasir-SEG: a segmented polyp dataset. In: Proceedings of the international conference on multimedia modeling (MMM)
21. Angermann Q, Bernal J, Sánchez-Montes C, Hammami M, Fernández-Esparrach G, Dray X, Romain O, Sánchez FJ, Histace A (2017) Real-time polyp detection in colonoscopy videos: a preliminary study for adapting still frame-based methodology to video sequences analysis. In: International journal of computer assisted radiology and surgery. https://hal.archives-ouvertes.fr/hal-01488657
22. Bernal J, Histace A, Masana M, Angermann Q, Sánchez-Montes C, Rodríguez de Miguel C, Hammami M, García-Rodríguez A, Córdova H, Romain O, Fernández-Esparrach G, Dray X, Sánchez FJ (2019) GTCreator: a flexible annotation tool for image-based datasets. Int J Comput Assist Radiol Surg 14:191–201. https://doi.org/10.1007/s11548-018-1864-x
23. Yu J, Pan H, Yin Q, Bian X, Cui Q (2019) Fully convolutional densenets for polyp segmentation in colonoscopy. In: 2019 IEEE 35th international conference on data engineering workshops (ICDEW), Macao, Macao, 2019. pp 306–311. https://doi.org/https://doi.org/10.1109/ICDEW.2019.00010
24. Shin Y, Qadir HA, Aabakken L, Bergsland J, Balasingham I (2018) Automatic colon polyp detection using region based deep CNN and post learning approaches. IEEE Access 6:40950–40962. https://doi.org/10.1109/ACCESS.2018.2856402
25. Bernal J, Sánchez FJ, Rodríguez de Miguel C, Fernández-Esparrach G (2015) Building up the future of colonoscopy? A synergy between clinicians and computer scientists. In: Screening for colorectal cancer with colonoscopy. InTech
26. Guo X, Zhang N, Guo J, Zhang H, Hao Y, Hang J (2019) Automated polyp segmentation for colonoscopy images: a method based on convolutional neural networks and ensemble learning. Med Phys 46:5666–5676. https://doi.org/10.1002/mp.13865
27. Kang J, Gwak J (2019) Ensemble of instance segmentation models for polyp segmentation in colonoscopy images. IEEE Access 7:26440–26447. https://doi.org/10.1109/ACCESS.2019.2900672
28. Akbari M, Mohrekesh M, Nasr-Esfahani E, Soroushmehr SMR, Karimi N, Samavi S, Najarian K (2018) Polyp segmentation in colonoscopy images using fully convolutional network. arXiv. https://arxiv.org/abs/1802.00368
29. Brandao P, Zisimopoulos O, Mazomenos E, Ciuti G, Bernal J, Visentini-Scarzanella M, Menciassi A, Dario P, Koulaouzidis A, Arezzo A, Hawkes DJ, Stoyanov D (2018) Towards a computed-aided diagnosis system in colonoscopy: automatic polyp segmentation using convolution neural networks. J Med Robot Res 3:1–13. https://doi.org/10.1142/S2424905X18400020
30. Wichakam I, Panboonyuen T, Udomcharoenchaikit C (2018) Real-time polyps segmentation for colonoscopy video frames using compressed fully convolutional network. In: Multimedia modeling. MMM 2018. Lecture notes in computer science, vol 10704. pp 393–404
31. Wickstrøm K, Kampffmeyer M, Jenssen R (2018) Uncertainty modeling and interpretability in convolutional neural networks for polyp segmentation. In: 2018 IEEE international workshop on machine learning for signal processing
32. Bardhi O, Member I, Sierra-Sosa D, Garcia-Zapirain B, Elmaghraby A (2017) Automatic colon polyp detection using Convolutional Encoder-Decoder model. In: 2017 IEEE international symposium on signal processing and information technol-

ogy (ISSPIT). pp 445–448. https://doi.org/https://doi.org/10.1109/ISSPIT.2017.8388684
33. Li Q, Yang G, Chen Z, Huang B, Chen L, Xu D, Zhou X, Zhong, Zhang H, Wang T (2017) Colorectal polyp segmentation using a fully convolutional neural network. In: 2017 10th international congress on image and signal processing, biomedical engineering and informatics (CISP-BMEI). pp 1–5. https://doi.org/https://doi.org/10.1109/CISP-BMEI.2017.8301980
34. Abaza A, Harrison MA, Bourlai T (2012) Quality metrics for practical face recognition. In Proceedings of the proceedings of the 21st international conference on pattern recognition (ICPR2012). pp 3103–3107
35. Tripathi AK, Mukhopadhyay S, Dhara AK (2011) Performance metrics for image contrast. In Proceedings of the 2011 international conference on image information processing. https://doi.org/https://doi.org/10.1109/ICIIP.2011.6108900
36. Ronneberger O, Fischer P, Brox T (2015) U-Net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells WM, Frangi AF (eds) Medical image computing and computer-assisted intervention—MICCAI 2015. Lecture notes in computer science, vol 9351. Springer, pp 234–241
37. Chollet F (2015) Keras. https://github.com/keras-team/keras
38. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Corrado GS, Davis A, Dean J, Devin M, Ghemawat S, Goodfellow I, Harp A, Irving G, Isard M, Jia Y, Jozefowicz R, Kaiser L, Kudlur M, Levenberg J, Mané D, Monga R, Moore S, Murray D, Olah C, Schuster M, Shlens J, Steiner B, Sutskever I, Talwar K, Tucker P, Vanhoucke V, Vasudevan V, Viégas F, Vinyals O, Warden P, Wattenberg M, Wicke M, Yu Y, Zheng X (2015) TensorFlow: large-scale machine learning on heterogeneous distributed systems. https://arxiv.org/abs/1603.04467. Accessed 05 Jan 2020
39. Shvets AA, Rakhlin A, Kalinin AA, Iglovikov VI (2018) Automatic instrument segmentation in robot-assisted surgery using deep learning. In: 2018 17th IEEE international conference on machine learning and applications ICMLA 2018. pp 624–628
40. Wilcox RR (2003) Comparing two independent groups. In: Wilcox RR (ed) Applying Contemporary Statistical Techniques. Elsevier Science, Berlin, pp 237–284
41. De Almeida Thomaz V, Sierra-Franco CA, Raposo AB (2019) Training data enhancements for robust polyp segmentation in colonoscopy images. In: Proceedings—IEEE symposium on computer-based medical systems. pp 192–197
42. Hwang M, Wang D, Jiang WC, Pan X, Fu D, Hwang KS, Ding K (2019) An adaptive regularization approach to colonoscopic polyp detection using a cascaded structure of encoder–decoders. Int J Fuzzy Syst 21:2091–2101. https://doi.org/10.1007/s40815-019-00694-y
43. Ross T, Zimmerer D, Vemuri A, Isensee F, Wiesenfarth M, Bodenstedt S, Both F, Kessler P, Wagner M, Müller B, Kenngott H, Speidel S, Kopp-Schneider A, Maier-Hein K, Maier-Hein L (2018) Exploiting the potential of unlabeled endoscopic video data with self-supervised learning. Int J Comput Assist Radiol Surg 13:925–933. https://doi.org/10.1007/s11548-018-1772-0
44. Zhang C, Tavanapong W, Kijkul G, Wong J, de Groen PC, Oh JH (2018) Similarity-based active learning for image classification under class imbalance. In: Proceedings of the IEEE international conference on data mining, ICDM 2018-November. pp 1422–1427. https://doi.org/https://doi.org/10.1109/ICDM.2018.00196
45. Tran T, Pham T, Carneiro G, Palmer L, Reid I (2017) A Bayesian data augmentation approach for learning deep models. In: 31st conference on neural information processing systems (NIPS 2017)
46. Saltiel D, Benhamou E (2018) Feature selection with optimal coordinate ascent (OCA). SSRN Electron J. https://doi.org/10.2139/ssrn.3293503
47. Zarshenas A, Suzuki K (2016) Binary coordinate ascent: an efficient optimization technique for feature subset selection for machine learning. Knowl Based Syst 110:191–201. https://doi.org/10.1016/j.knosys.2016.07.026

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

8.4 Eigenloss: combined PCA-based loss function for polyp segmentation

Dr. J. Blas Pagador Carrasco y Dr. Artzai Picón Ruiz, como directores de la tesis titulada "Mejoras en la detección, localización y segmentación automática de pólipos colorrectales basadas en técnicas de aprendizaje profundo", certificamos el factor de impacto y la categorización de la siguiente publicación, incluida en la tesis doctoral. Del mismo modo, se especifica la aportación del doctorando.

Autores: **Luisa F. Sánchez-Peralta**, Artzai Picón, Juan Antonio Antequera-Barroso, Juan Francisco Ortega-Morán, Francisco M. Sánchez-Margallo, J. Blas Pagador

Título: Eigenloss: Combined PCA-Based Loss Function for Polyp Segmentation

Revista: Mathematics – Número especial: Recent Advances in Deep Learning

Otros datos: vol 8, nº 8, pág. 1316, año 2020

DOI: 10.3390/math8081316

Factor de impacto (2020): 2.258 (Q1 – Mathematics)

Abstract: Colorectal cancer is one of the leading cancer death causes worldwide, but its early diagnosis highly improves the survival rates. The success of deep learning has also benefited this clinical field. When training a deep learning model, it is optimized based on the selected loss function. In this work, we consider two networks (U-Net and LinkNet) and two backbones (VGG-16 and Densnet121). We analyzed the influence of seven loss functions and used a principal component analysis (PCA) to determine whether the PCA-based decomposition allows for the defining of the coefficients of a non-redundant primal loss function that can outperform the individual loss functions and different linear combinations. The eigenloss is defined as a linear combination of the individual losses using the elements of the eigenvector as coefficients. Empirical results show that the proposed eigenloss improves the general performance of individual loss functions and outperforms other linear combinations when Linknet is used, showing potential for its application in polyp segmentation problems.

Contribución del doctorando: Conceptualization, methodology, software, validation, formal analysis, investigation, data curation, writing—original draft preparation, writing—review and editing, visualization.

Dr. D. José Blas Pagador Carrasco Dr. D. Artzai Picón Ruiz

© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).

mathematics

MDPI

Article

Eigenloss: Combined PCA-Based Loss Function for Polyp Segmentation

Luisa F. Sánchez-Peralta [1,*], Artzai Picón [2], Juan Antonio Antequera-Barroso [3], Juan Francisco Ortega-Morán [1], Francisco M. Sánchez-Margallo [1] and J. Blas Pagador [1]

1 Jesús Usón Minimally Invasive Surgery Centre, N-521, km 41.7, E-10071 Cáceres, Spain; jfortega@ccmijesususon.com (J.F.O.-M.); msanchez@ccmijesususon.com (F.M.S.-M.); jbpagador@ccmijesususon.com (J.B.P.)
2 TECNALIA, Basque Research and Technology Alliance (BRTA), Parque Tecnológico de Bizkaia, C/Geldo. Edificio 700, E-48160 Derio, Spain; artzai.picon@tecnalia.com
3 Didactics of Mathematics, University of Cadiz, Avda. República Saharaui s/n. Campus de Puerto Real, E-11519 Puerto Real, Spain; juanantonio.antequera@uca.es
* Correspondence: lfsanchez@ccmijesususon.com; Tel.: +34-927-18-10-32

Received: 8 June 2020; Accepted: 5 August 2020; Published: 7 August 2020

check for updates

Abstract: Colorectal cancer is one of the leading cancer death causes worldwide, but its early diagnosis highly improves the survival rates. The success of deep learning has also benefited this clinical field. When training a deep learning model, it is optimized based on the selected loss function. In this work, we consider two networks (U-Net and LinkNet) and two backbones (VGG-16 and Densnet121). We analyzed the influence of seven loss functions and used a principal component analysis (PCA) to determine whether the PCA-based decomposition allows for the defining of the coefficients of a non-redundant primal loss function that can outperform the individual loss functions and different linear combinations. The eigenloss is defined as a linear combination of the individual losses using the elements of the eigenvector as coefficients. Empirical results show that the proposed eigenloss improves the general performance of individual loss functions and outperforms other linear combinations when Linknet is used, showing potential for its application in polyp segmentation problems.

Keywords: deep learning; loss functions; principal component analysis; polyp segmentation

1. Introduction

Nowadays, colorectal cancer (CRC) presents high incidence rates in developed countries, as it is mainly considered a lifestyle disease [1]. More than 800,000 deaths are reported yearly and CRC incidence reaches one in ten people worldwide [2]. In 2020, over 104,000 new cases and 53,200 deaths are estimated in the United States alone [3], while more than 174,600 deaths are foreseen in Europe [4]. However, an early detection of CRC, increasing the current 13% up to 50% of patients diagnosed in stage I could save 130,000 lives and reduce the healthcare cost per year by three billion [5]. In this sense, colonoscopy is the gold standard for early diagnosis (CRC screening): the gastroenterologist performs a visual exploration of the rectum and colon using a flexible endoscope inserted through the patient's anus. During the procedure, polyps are detected and treated if deemed necessary, as these are considered precursor lesions of CRC. Colonoscopy is highly dependent on the endoscopist's skills, therefore computer-aided diagnosis and detection systems have a great potential to improve clinical results [6].

Recently, deep learning has outperformed classical methods on several computer vision tasks, such as detection, segmentation, classification, and enhancement, also in the healthcare field [7]. The

automatic extraction of features performed by convolutional neural networks (CNN) is the main advantage over classical hand-crafted features, but the need of large and annotated datasets is one of its major drawbacks in specific fields, such as medical applications. For this reason, several studies focus on improving the scarce and weak annotations in medical imaging datasets [8]. Although clinical evidence of deep learning algorithms is still poor (there are few studies with low confidence), they show equivalent (or even better) performance than clinicians [9,10]. In this sense, some systematic reviews show the potential of deep learning in gastric tissue diseases [11] and wireless endoscopic capsule [12], but these studies also identify risk of bias due to gaps in the evaluation metrics and public availability of the dataset [11] that must be solved through prospective multicenter studies [12].

In our field of interest, computer-aided detection (CAD) systems have been classically based on the extraction of polyp characteristics, like shape [13], texture [14,15] or depth of valleys [16,17]. Moreover, other classical segmentation methods could be also applied to this problem, such as the alternating direction method of multipliers (ADMM) approach to decompose the image into non-additive components and classify pixels into background and foreground [18] or the approach proposed by Li et al. [19], who combine semi-supervised learning of the posterior class distribution modeled using multinomial logistic regression with segmentation that exploits spatial information. It has been recently determined that up to 80% of missed colonic lesions could be avoided using real-time artificial intelligence assistance [20]. Therefore, there is still a need for further developing polyp detection, localization, and segmentation methods. In this last case, several authors use deep learning methods, which outperform hand-crafted methods [21], and where CNNs are widely used [22]. Among them, the encoder-decoder architectures are a popular family of deep models for semantic segmentation [23]. U-Net [24] is one of the most widely used and well-known architectures for medical imaging segmentation, which has led to many variants for particular applications with great results [25–28]. While U-Net relies on concatenation in the decoder, LinkNet [29] adds the information from the encoder and has also been used for medical image segmentation [30,31] including polyp segmentation [32]. Furthermore, these two networks are usually compared in studies from different medical fields [33–36]. For polyp segmentation on colonoscopy images, different approaches can be found. Wichakam et al. [37] proposed a compressed fully convolutional network (FCN) based on VGG-16 [38] which provided state of the art results (Jaccard index of 69.36) while running seven times faster than the original FCN. Vázquez et al. [39] employ a FCN8 architecture [40] and get a Jaccard index equal to 56.07 for the polyp class. Similarly, Wickstrøm et al. [41], also with an enhanced FCN8, get 58.70 overlap.

Principal Component Analysis (PCA) is a mathematical technique for data transformation that reduces multidimensional data sets into a lower number of principal components, which are uncorrelated and retain variance as much as possible [42]. PCA has been used in medical imaging with different purposes. Ansari et al. [43] use the PCA to obtain a target image from a different source image. They transform endoscopic narrow-band images (NBI) into standard colored endoscopic images. In this case, the PCA extracts the fundamental information in NBI, which usually enhances structures and textures, so the standard images can be improved for better assessment and diagnosis. Another application of PCA is the reduction of highly dimensional data, such as fluorescence spectral images of colorectal polyps [44]. In this work, near-infrared autofluorescence spectroscopy images with multiple wavelengths and intensities were obtained from healthy, hyperplastic, and adenomatous colonic tissues and this high dimensional set of images was reduced to eight principal components for an easier and more intuitive tissue classification. It is also worth mentioning that other technical aspects, such as prevention of overfitting in neural networks, have also been addressed with PCA. Kim et al. [45] apply the PCA to reduce the dimensions of the input feature vector while retaining the main information, so the number of weights in the neural network can be reduced. Related to polyp segmentation, PCA has been used as reduction technique for contour region analysis based on hand-crafted features [46].

In statistics, a loss function is commonly used in optimization problems, where the minimum difference between estimated and true values for a given dataset is sought. Therefore, loss functions,

also known as objective functions, cost functions, error functions or simply losses, conduct the training process of deep learning models computing the error between predicted values and ground truth and minimizing it to optimize the weights of the model. The relevance and influence of loss functions in deep learning models have been studied in many fields, such as handwritten digits recognition [47], fast single classification [48], cardiac MRI reconstruction [49], low-dose CT images denoising [50], or face recognition [51]. Loss functions can be categorized into distribution, region or boundary-based losses, depending on the main concept considered for minimization. Regarding distribution-based losses, the cross entropy loss function [52] is widely used. It calculates the cost at each pixel using a logarithmic function that is then averaged over all pixels to quantify the difference between the two probability distributions (ground truth vs prediction). Other distributed-based functions are the weight cross entropy [24] and the focal loss [53]. In case of region-based losses, the commonly used Dice loss [25] measures the overlap of two sets, which can be easily applied to binary segmentation. Additional region-based functions are the Jaccard or Intersection over Union (IoU) loss [54], the Tversky loss [55], the Lovász-Hinge or Lovasz-Softmax loss [56], and the focal Tversky loss [57]. Finally, boundary-based loss functions include for example the surface loss that uses a distance metric to compare the shapes of two contours [58], or the loss that uses the estimations of the Hausdorff distance to measure differences between two contours [59].

When deep learning is applied to medical imaging analysis, researchers tend to use classical cross-entropy loss functions together with a second distance or overlap measure [60], so combined loss functions are commonly used [61–63]. For polyp segmentation, both individual and combined loss functions can be found. Wichamkam et al. [37] apply the Dice loss function alone. On the other hand, while Zhou [26] combines the Dice loss function with the binary cross entropy, Mohammed et al. [64] combine it with the weight binary cross entropy instead.

Hence, the aim of this paper is to analyze the influence of several loss functions on different models for polyp segmentation and to determine whether a PCA-based decomposition allows for the defining of the coefficients of a non-redundant primal loss function that can outperform the individual loss functions and other aggregation methods, such as the sum and the mean of loss functions. Consequently, the same dataset and training conditions were used for all experiments, while several loss functions were applied to compare the training and testing performance of the model. Two different encoder-decoder architectures, namely U-Net [24] and LinkNet [29], with two backbones each (VGG-16 [40] and Densenet121 [65]) were used for comparison.

The rest of the paper is organized as follows. In Section 2, a detailed description of materials and methods is exposed, including dataset, network architectures, training parameters, loss functions, metrics and PCA analysis. In Section 3, the results are presented and analyzed. Finally, Section 4 discusses the results and Section 5 summarizes the most important conclusions of this work.

2. Materials and Methods

2.1. Dataset

CVC-EndoSceneStill is a publicly available dataset [39] that contains 912 images obtained from 44 video sequences collected from 36 patients. The dataset is divided into training, validation and test sets, each of them including 547, 183, and 182 images, respectively. For each image, binary masks are provided for the polyp, lumen, specular lights and void (black area) classes. In the binary masks, pixels corresponding to the class are labelled with 1, and 0 otherwise (Figure 1). In this work, only the polyp and void classes are used. While the polyp binary mask is used both at training and testing, the void binary mask is just used for reporting metrics with respect to the valid endoscopic image.

(**a**) (**b**) (**c**)

Figure 1. Example of the CVC-EndoSceneStill dataset (**a**) polyp image; (**b**) polyp binary mask; (**c**) void binary mask.

2.2. Architectures and Training Parameters

For this study, two backbones (VGG-16 [40] and Densenet121 [65]) and two encoder-decoder architectures (U-Net [24] and LinkNet [29]) were selected. The backbones differ in the type of block used to obtain the feature vector that is fed into the decoder. VGG-16 (Figure 2A) employs convolutional blocks composed of convolutional and activation layers, which are then followed by a max pooling layer to reduce the input size. On the other hand, Densenet121 (Figure 2B) relies on dense blocks where all layers are interconnected with the rest of layers in the block. In both cases, backbones are pretrained using IMAGENET [66], rather than trained from scratch, in order to improve the performance on a small dataset such as CVC-EndosceneStill (only 547 images for training). As for the encoder-decoder architectures, the difference between U-Net and LinkNet mainly lies on the integration of the information of the encoder into the decoder. While U-Net (Figure 2C) concatenates the features maps of the encoder at their corresponding level in the decoder, LinkNet (Figure 2D) adds the input to each encoder layer to the output of the corresponding layer in the decoder. In our implementation, the decoder path is created accordingly to the selected backbone to match the dimensions and number of layers. Therefore, we have four different models (Tables S1–S4 provided in the Supplementary Materials): U-Net-VGG-16 (Figure 2A,C), U-Net-Densenet121 (Figure 2B,C), LinkNet-VGG-16 (Figure 2A,D) and LinkNet-Densenet121 (Figure 2B,D). In all cases, image input size is set to $256 \times 256 \times 3$ and the predicted binary mask size is $256 \times 256 \times 1$.

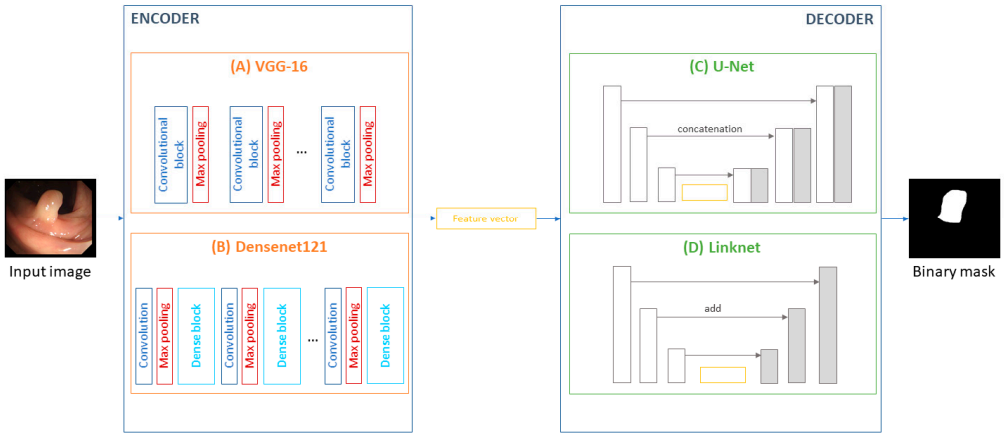


Figure 2. Backbones for the encoder (**A**,**B**) and encoder-decoder architectures (**C**,**D**).

Each model is trained for 300 epochs using the Adam optimizer, a batch size of 16 and batch normalization. The learning rate is initially set to 0.001, decreasing by half every 5 epochs and restoring the initial value every 25 epochs. The weights of the best epoch (lowest validation loss) are used

for prediction. All experiments were implemented using segmentation models [67], Keras [68] and Tensorflow [69] as backend and were run in a NVIDIA Quadro P5000 with 16 GB memory.

2.3. Loss functions

Seven loss functions are considered in this work. In order to present their formulation, we firstly establish the mathematical framework of the binary segmentation problem. Since we are performing a binary semantic segmentation based on a pixel-wise classification, we describe the problem of a binary classifier.

As $\boldsymbol{X}$ is the image of size M × N to be segmented, each pixel $x \in \boldsymbol{X}$ has a corresponding value, y , in the ground truth mask, $\boldsymbol{Y}$. In a binary classification problem, the ground truth labels can take the values $y = \{y_1, y_0\}$, where y_1 is the positive class, being polyp in this case, and y_0 the negative class, corresponding to the background class. The segmentation model gives a prediction $\hat{\boldsymbol{Y}}$, with labels $\hat{y} = \{\hat{y}_1, \hat{y}_0\}$. In our work, labels are assigned as $y_1 = 1;\ y_0 = 0;\ \hat{y}_1 = 1;\ \hat{y}_0 = 0$. Thus, the elements of the confusion matrix, n_{ij}, indicate the number of cases belonging to class i that were classified as class j. In our binary problem, the elements of the 2 × 2 confusion matrix are:

- n_{11}, or true positives (TP): polyp pixels in the ground truth binary mask that are correctly classified as polyp in the predicted binary mask.
- n_{10}, or false negatives (FN): polyp pixels in the ground truth binary mask that are incorrectly classified as background in the predicted binary mask.
- n_{01}, or false positives (FP): background pixels in the ground truth binary mask that are incorrectly classified as polyp in the predicted binary mask.
- n_{00}, or true negatives (TN): background pixels in the ground truth binary mask that are correctly classified as background in the predicted binary mask.

Therefore, any loss function $L\left(\boldsymbol{Y}, \hat{\boldsymbol{Y}}\right)$ measures the cost or error of predicting $\hat{\boldsymbol{Y}}$ for a given image $\boldsymbol{X}$ with ground truth $\boldsymbol{Y}$ [70]. In the following subsections, we define the loss functions considered in this study.

2.3.1. Jaccard Loss

The Jaccard index [71], also known as intersection over union (IoU), provides the ratio of the intersection between the ground truth and the predicted binary masks over their union, in a [0, 1] range and with the convention of solving the indetermination $0/0 = 1$. In our particular dataset, this situation does not occur as all images contains a segmented polyp, so the ground truth binary masks always contain positive elements. Based on this index, the Jaccard loss can be defined as:

$$L_{Jaccard}\left(\boldsymbol{Y}, \hat{\boldsymbol{Y}}\right) = 1 - \frac{\left|\boldsymbol{Y} \cap \hat{\boldsymbol{Y}}\right|}{\left|\boldsymbol{Y} \cup \hat{\boldsymbol{Y}}\right|} = 1 - \frac{\sum_{i=1}^{M \cdot N} y_i \cdot \hat{y}_i}{\sum_{i=1}^{M \cdot N} y_i + \sum_{i=1}^{M \cdot N} \hat{y}_i - \sum_{i=1}^{M \cdot N} y_i \cdot \hat{y}_i} = 1 - \frac{TP}{TP + FN + FP} \tag{1}$$

This loss function is widely used in segmentation problems as it directly optimizes the Jaccard index, one of the most common metrics.

2.3.2. Dice Loss

The Dice coefficient [72] also measures similarities between two regions. This coefficient is the same as the F1-score. The Dice loss can be defined as:

$$L_{Dice}\left(\boldsymbol{Y}, \hat{\boldsymbol{Y}}\right) = 1 - \frac{\mathbf{2} \cdot \left|\boldsymbol{Y} \cap \hat{\boldsymbol{Y}}\right|}{|\boldsymbol{Y}| + \left|\hat{\boldsymbol{Y}}\right|} = 1 - \frac{2 \sum_{i=1}^{M \cdot N} y_i \cdot \hat{y}_i}{\sum_{i=1}^{M \cdot N} y_i + \sum_{i=1}^{M \cdot N} \hat{y}_i} = 1 - \frac{2 \cdot TP}{2 \cdot TP + FN + FP} \tag{2}$$

2.3.3. Binary Cross Entropy Loss

Cross entropy or softmax loss is the most widely used loss function for classification problems [73]. For a binary classification problem, binary cross entropy can be defined as:

$$L_{BEC}\left(\boldsymbol{Y}, \hat{\boldsymbol{Y}}\right) = \frac{1}{M{\cdot}N} \sum_{\mathrm{i}=1}^{M{\cdot}N} \left(-y_i{\cdot}\log(\hat{y}_i) - (1-y_i){\cdot}\log(1-\hat{y}_i)\right) \tag{3}$$

The aim of this distributed-based loss function is to maximize the accuracy of the predicted binary mask, considering both the positive class, or polyp, and the negative class, or background.

2.3.4. Binary Focal Loss

Yin et al. [53] defined the focal loss to counteract the effect of imbalanced classes. For a binary problem, it can be defined as:

$$L_{BF}\left(\boldsymbol{Y}, \hat{\boldsymbol{Y}}\right) = \sum_{\mathrm{i}=1}^{M{\cdot}N} \left(-y_i{\cdot} \propto {\cdot}(1-\hat{y}_i)^{\gamma}{\cdot}\log(\hat{y}_i) - (1-y_i){\cdot} \propto {\cdot}\hat{y}_i{}^{\gamma}{\cdot}log(1-\hat{y}_i)\right) \tag{4}$$

being α a weighting factor in balanced cross entropy and γ a focusing parameter. We have considered $\alpha = 0.25$ and $\gamma = 2.0$, default parameters in the segmentation models library [67].

2.3.5. Tversky Loss

The Tversky index [74] allows us to control the penalties for FP and FN through the parameters α and β, and it is defined as:

$$TI\left(\boldsymbol{Y}, \hat{\boldsymbol{Y}}\right) = \frac{\sum_{i=1}^{M{\cdot}N} y_{pi}\hat{y}_{pi} + \epsilon}{\sum_{i=1}^{M{\cdot}N} y_{pi}\hat{y}_{pi} + \alpha \sum_{i=1}^{M{\cdot}N} y_{ip}\hat{y}_{\overline{ip}} + \beta \sum_{i=1}^{M{\cdot}N} y_{\overline{ip}}\hat{y}_{ip} + \epsilon} \tag{5}$$

where p stands for the polyp class, so $y_{pi} = 1$ if pixel i belongs to the polyp class, and 0 otherwise, and the opposite for $y_{\overline{p}i}$ in the ground truth binary mask. The same definitions apply for $\hat{y}_{pi}$ and $\hat{y}_{\overline{p}i}$ in the predicted binary mask. Based on this index, the Tversky loss function is defined as:

$$L_{Tversky}\left(\boldsymbol{Y}, \hat{\boldsymbol{Y}}\right) = 1 - TI\left(\boldsymbol{Y}, \hat{\boldsymbol{Y}}\right) \tag{6}$$

For implementation, we have selected $\alpha = 0.7$ and $\beta = 1 - \alpha = 0.3$, as the authors report in their original work.

2.3.6. Focal Tversky Loss

Based on the Tversky index (Equation (6)) and combining the idea of the binary focal loss (Section 2.3.4), Abraham et al. [57] defined the Focal Tversky loss as:

$$L_{FocalTversky}\left(\boldsymbol{Y}, \hat{\boldsymbol{Y}}\right) = \left(1 - TI\left(\boldsymbol{Y}, \hat{\boldsymbol{Y}}\right)\right)^{\frac{1}{\gamma}} \tag{7}$$

to address the issue of data imbalance in medical imaging segmentation. For implementation, we have selected $\gamma = \frac{4}{3}$, as the authors report in their original work.

2.3.7. Lovász-Hinge Loss

Berman et al. [56] propose the Lovász-Hinge loss function, based on the convex Lovász extension of sub-modular losses. For our binary problem, it is defined as:

$$L_{Lovasz-Hinge}\left(\boldsymbol{Y}, \hat{\boldsymbol{Y}}\right) = \overline{\Delta_{J_p}}(\boldsymbol{m}(p)) \tag{8}$$

where $\overline{\Delta_{J_p}}$ is the Lovász extension to the Jaccard loss function and $\boldsymbol{m}(p)$ is the vector of hinge losses for the polyp class, as defined in [56].

2.4. Metrics

The results are always reported over the 182 images of the test set of CVC-EndoSceneStill dataset, which allows for a fair comparison of the models. To calculate the reporting metrics, only the valid area of the void binary mask (Figure 1c) is considered. For each image, seven metrics are calculated based on the elements of the confusion matrix:

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

$$Precision = \frac{TP}{TP + FP} \tag{10}$$

$$Recall = \frac{TP}{TP + FN} \tag{11}$$

$$Specificity = \frac{TN}{TN + FP} \tag{12}$$

$$F2 - score = \frac{5 \cdot TP}{5 \cdot TP + 4 \cdot FN + FP} \tag{13}$$

$$Jaccard = \frac{TP}{TP + FN + FP} \tag{14}$$

$$Dice = \frac{2 \cdot TP}{2 \cdot TP + FN + FP} \tag{15}$$

2.5. PCA Analysis

Previously to the application of the PCA, the viability of the method was tested. Therefore, the existing correlation between the variables under study was determined based on the correlation matrix and two statistical tests: Kaiser-Meyer-Olkin (KMO) index and χ^2 test. Next, the Bartlett sphericity test was carried out to determine the correlation between all variables. This test is based on Pearson's χ^2 distribution, indicating that a significant correlation between variables other than zero when the determinant of the correlation matrix is different from one. The mean of the KMO sample fit was also calculated to measure the proportion of variance that the variables under study have in common. This parameter ranges from 0 to 1, considering a good sample suitability to use PCA for values greater than 0.5.

Under the assumption that these tests are positive, a PCA was performed using the seven metrics of the 182 images of the test set predicted by the model trained with each individual loss function, in order to identify non-correlated variables that allow for a dimensionality reduction while the included variance is maximized. We have a set of data $X \in \mathbb{R}^{N \times D}$, where N is the number of metrics calculated for the 182 images of the test set and D is the number of loss functions. The objective is to project X into Y$\in \mathbb{R}^{N \times K}$, where $K < D$, while maximizing the variance included in Y. If the D variables are correlated, the information in X is redundant, so the PCA analysis allows for a dimension reduction resulting in a set of independent variables. For that, eigenvectors and eigenvalues from the correlation matrix are calculated. Eigenvalues are sorted in descending order and K eigenvectors with the largest eigenvalues are selected. Then, the original dataset X can be transformed into Y using the projection matrix $\boldsymbol{W}$ from the selected K eigenvectors as:

$$Y = \boldsymbol{W}X = \alpha_{k1}X_1 + \alpha_{k2}X_2 + \ldots + \alpha_{KN}X_N \tag{16}$$

being α_{kn} the coefficient of the *k-th* principal component for the *n-th* input variable. The resulting components were examined and only those with eigenvalues greater than 1 which also include a great

proportion of variance (over 80%) were considered [75]. We can define an eigenloss function for each principal component as a linear combination of the individual loss functions as:

$$\begin{aligned} L_{k\ Eigenloss}\left(\boldsymbol{Y},\hat{\boldsymbol{Y}}\right) = \quad & \alpha_{k\ Jaccard}\cdot L_{Jaccard}\left(\boldsymbol{Y},\hat{\boldsymbol{Y}}\right) + \alpha_{k\ Dice}\cdot L_{Dice}\left(\boldsymbol{Y},\hat{\boldsymbol{Y}}\right) + \alpha_{k\ BCE} \\ & \cdot L_{BCE}\left(\boldsymbol{Y},\hat{\boldsymbol{Y}}\right) + \alpha_{k\ BF}\cdot L_{BF}\left(\boldsymbol{Y},\hat{\boldsymbol{Y}}\right) + \alpha_{k\ Tversky}\cdot L_{Tversky}\left(\boldsymbol{Y},\hat{\boldsymbol{Y}}\right) \\ & + \alpha_{k\ FocalTversky}\cdot L_{FocalTversky}\left(\boldsymbol{Y},\hat{\boldsymbol{Y}}\right) + \alpha_{k\ Lovasz\ Hinge} \\ & \cdot L_{Lovasz\ Hinge}\left(\boldsymbol{Y},\hat{\boldsymbol{Y}}\right) \end{aligned} \tag{17}$$

3. Results and Analysis

3.1. PCA Analysis

Tables 1–4 summarize the results of the different models in the test set of CVC-EndoSceneStill. Metrics calculated in the [0, 1] range were transformed into [0, 100] range for better visualization. In all models, the accuracy is always closer to 100 due to the influence of the background in this metric and the unbalance between polyp and background classes. This is also found in the case of precision and specificity, where values are in general over 80.00 and 98.00, respectively. On the contrary, recall, F2-score, Jaccard and Dice provide lower values, ranging from around 40.00 to 81.00. It is important to highlight the small deviation of the specificity in comparison to the rest of the metrics. This is due to the high presence of background pixels, which makes the TN a very stable figure. When classes are highly unbalance, such as our case because the polyp represents 12.50 ± 11.49% (range: 0.75–66.15) of the image, metrics that do not consider the TN (precision, recall, F2-score, Jaccard and Dice) are more representative of the performance.

Table 1. U-Net-VGG-16. Results of the model trained with the different loss functions, reported over the test set of CVC-EndoSceneStill. Best values are indicated in bold.

Loss Function	Accuracy	Precision	Recall	Specificity	F2-Score	Jaccard	Dice
Jaccard	**93.36 ± 10.26**	**88.69 ± 23.02**	65.65 ± 37.52	99.24 ± 1.73	65.79 ± 36.20	59.77 ± 35.03	67.39 ± 34.44
Dice	92.76 ± 10.45	82.20 ± 27.36	63.24 ± 37.94	99.07 ± 1.66	63.23 ± 36.42	56.38 ± 34.59	64.61 ± 34.47
Binary entropy	93.18 ± 10.28	85.39 ± 30.95	64.62 ± 38.92	99.29 ± 1.45	64.89 ± 37.90	59.48 ± 36.35	66.39 ± 36.58
Binary focal	93.34 ± 10.28	82.44 ± 26.45	68.84 ± 36.37	98.59 ± 2.84	68.13 ± 34.91	60.49 ± 33.66	68.48 ± 33.59
Tversky	92.80 ± 10.53	82.80 ± 29.18	60.84 ± 39.16	99.22 ± 1.57	61.20 ± 38.03	55.34 ± 36.12	62.88 ± 36.59
Focal Tversky	93.10 ± 9.94	81.19 ± 26.54	64.03 ± 38.72	98.98 ± 2.13	63.81 ± 37.25	57.20 ± 35.57	64.89 ± 35.46
Lovász-Hinge	91.26 ± 11.07	72.26 ± 41.71	40.09 ± 39.55	**99.71 ± 0.86**	41.48 ± 39.44	38.23 ± 37.64	44.74 ± 39.49
Sum	92.78 ± 10.45	81.15 ± 32.97	57.76 ± 39.88	99.37 ± 1.44	58.43 ± 39.01	53.58 ± 37.44	60.64 ± 38.11
Mean	93.33 ± 10.5	84.07 ± 29.03	**69.53 ± 35.64**	98.84 ± 2.42	**69.38 ± 34.42**	**62.83 ± 33.45**	**70.49 ± 33.13**
Eigenloss	93.00 ± 10.91	80.31 ± 31.35	65.07 ± 37.70	99.00 ± 2.25	65.45 ± 36.70	59.66 ± 35.32	67.04 ± 35.39
Norm Eigenloss	92.97 ± 10.13	81.62 ± 27.02	64.02 ± 37.73	98.91 ± 2.06	63.82 ± 35.99	56.56 ± 33.99	65.01 ± 33.99

Table 2. U-Net-Densenet121. Results of the model trained with the different loss functions, reported over the test set of CVC-EndoSceneStill. Best values are indicated in bold.

Loss Function	Accuracy	Precision	Recall	Specificity	F2-Score	Jaccard	Dice
Jaccard	**95.02 ± 9.53**	86.95 ± 23.58	79.18 ± 33.08	**99.14 ± 1.66**	78.56 ± 31.94	72.17 ± 31.06	78.45 ± 30.37
Dice	94.17 ± 10.22	**90.10 ± 24.86**	77.88 ± 33.74	98.99 ± 2.34	76.63 ± 32.63	69.88 ± 32.63	76.30 ± 31.61
Binary entropy	94.30 ± 9.68	86.58 ± 23.06	75.60 ± 33.75	99.12 ± 1.65	74.96 ± 32.17	68.07 ± 31.89	75.39 ± 30.06
Binary focal	94.27 ± 9.81	84.61 ± 22.94	80.84 ± 31.62	98.64 ± 2.67	78.69 ± 30.03	70.01 ± 30.18	77.40 ± 28.41
Tversky	94.74 ± 9.77	89.04 ± 19.81	**81.04 ± 30.35**	99.02 ± 2.29	**79.71 ± 29.35**	73.19 ± 30.17	79.72 ± 27.98
Focal Tversky	94.19 ± 9.83	83.08 ± 25.12	78.59 ± 34.18	98.62 ± 2.90	68.05 ± 32.22	**75.10 ± 31.24**	**79.84 ± 24.73**
Lovász-Hinge	94.73 ± 9.77	87.52 ± 21.59	80.07 ± 31.72	98.95 ± 2.37	79.01 ± 30.42	72.26 ± 30.62	78.86 ± 28.78
Sum	94.76 ± 9.53	87.92 ± 22.08	80.72 ± 30.32	98.95 ± 2.38	79.69 ± 29.12	72.89 ± 29.77	79.62 ± 27.80
Mean	94.45 ± 9.77	86.12 ± 22.40	80.45 ± 31.49	98.82 ± 2.56	78.48 ± 30.17	70.75 ± 30.84	77.72 ± 28.96
Eigenloss	94.43 ± 9.78	89.83 ± 21.63	77.39 ± 33.28	99.06 ± 2.18	76.41 ± 31.97	69.35 ± 31.55	76.33 ± 30.39
Norm Eigenloss	94.68 ± 9.69	87.85 ± 22.12	79.24 ± 32.31	99.09 ± 1.95	78.43 ± 31.11	71.86 ± 31.02	78.37 ± 29.60

Table 3. LinkNet-VGG-16. Results of the model trained with the different loss functions, reported over the test set of CVC-EndoSceneStill. Best values are indicated in bold.

Loss Function	Accuracy	Precision	Recall	Specificity	F2-Score	Jaccard	Dice
Jaccard	93.41 ± 10.47	82.86 ± 30.16	68.2 ± 36.67	99.34 ± 1.22	68.51 ± 35.56	62.75 ± 34.19	70.07 ± 34.11
Dice	**93.59 ± 10.38**	81.03 ± 32.94	67.58 ± 38.01	**99.49 ± 0.96**	67.90 ± 37.10	62.53 ± 35.45	69.24 ± 35.92
Binary entropy	93.61 ± 9.98	83.01 ± 26.67	**70.18 ± 35.97**	98.97 ± 1.97	**69.76 ± 34.50**	62.65 ± 33.05	70.48 ± 32.88
Binary focal	93.15 ± 10.28	81.04 ± 28.02	66.18 ± 38.02	98.95 ± 1.76	65.63 ± 36.44	58.36 ± 34.85	66.22 ± 34.65
Tversky	93.45 ± 10.78	82.54 ± 28.10	69.82 ± 37.05	99.13 ± 1.85	69.55 ± 35.85	63.14 ± 34.49	70.27 ± 34.33
Focal Tversky	93.56 ± 10.35	83.56 ± 26.75	70.18 ± 37.39	99.09 ± 1.74	69.54 ± 36.10	62.81 ± 34.66	69.91 ± 34.58
Lovász-Hinge	92.98 ± 10.28	80.57 ± 29.84	68.17 ± 37.82	98.73 ± 2.60	67.21 ± 36.06	59.6 ± 34.49	67.42 ± 34.23
Sum	93.26 ± 10.08	78.00 ± 32.57	68.06 ± 36.90	98.76 ± 2.59	67.53 ± 35.50	60.4 ± 34.20	68.17 ± 34.12
Mean	93.53 ± 10.36	84.11 ± 28.24	68.64 ± 37.03	99.35 ± 1.23	68.80 ± 35.79	62.77 ± 34.16	70.10 ± 34.05
Eigenloss	93.50 ± 10.60	83.31 ± 28.49	69.52 ± 36.50	99.28 ± 1.76	69.51 ± 35.31	**63.24 ± 33.86**	**70.61 ± 33.68**
Norm Eigenloss	93.12 ± 10.59	**84.35 ± 28.65**	65.74 ± 37.70	99.33 ± 1.61	66.13 ± 36.53	60.62 ± 35.21	67.93 ± 35.03

Table 4. LinkNet-Densenet121. Results of the model trained with the different loss functions, reported over the test set of CVC-EndoSceneStill. Best values are indicated in bold.

Loss Function	Accuracy	Precision	Recall	Specificity	F2-Score	Jaccard	Dice
Jaccard	**94.71 ± 9.08**	87.32 ± 20.08	**81.92 ± 27.25**	98.58 ± 3.05	**79.84 ± 26.01**	71.57 ± 28.02	**79.44 ± 24.79**
Dice	94.47 ± 9.90	86.44 ± 23.22	76.91 ± 34.45	99.19 ± 1.52	76.27 ± 33.16	69.71 ± 32.18	76.32 ± 31.22
Binary entropy	93.99 ± 10.18	84.52 ± 25.45	77.44 ± 34.21	98.70 ± 2.73	75.96 ± 32.98	68.74 ± 32.73	75.44 ± 31.57
Binary focal	94.39 ± 9.88	85.15 ± 25.81	77.04 ± 34.42	99.08 ± 2.07	75.99 ± 33.15	69.11 ± 32.62	75.71 ± 31.76
Tversky	94.68 ± 9.41	88.13 ± 21.48	78.47 ± 31.38	99.00 ± 2.37	77.74 ± 29.93	70.92 ± 30.11	78.12 ± 28.08
Focal Tversky	94.31 ± 9.64	85.96 ± 21.48	81.01 ± 29.48	98.52 ± 3.11	79.16 ± 27.99	70.84 ± 28.83	78.53 ± 26.54
Lovász-Hinge	94.89 ± 9.56	86.73 ± 22.64	79.62 ± 31.09	**99.21 ± 1.29**	79.02 ± 30.08	**72.41 ± 29.88**	79.13 ± 28.59
Sum	94.36 ± 9.65	84.13 ± 25.73	76.50 ± 34.19	99.01 ± 1.67	75.47 ± 32.69	68.36 ± 32.11	75.45 ± 30.70
Mean	94.11 ± 10.09	85.48 ± 26.24	74.92 ± 34.48	98.96 ± 2.67	74.17 ± 33.11	67.77 ± 32.82	74.7 ± 31.61
Eigenloss	94.59 ± 9.61	**88.65 ± 18.25**	80.23 ± 30.78	99.05 ± 1.92	79.18 ± 29.26	71.84 ± 29.25	79.07 ± 27.07
NormEigenloss	94.60 ± 9.86	87.79 ± 21.65	77.81 ± 32.85	99.16 ± 1.89	77.49 ± 31.66	71.57 ± 31.28	78.06 ± 29.85

At loss function level, the Jaccard and Tversky losses appear to provide the most stable results along the different models. On the contrary, the binary cross entropy, the binary focal and the Lovász-Hinge loss functions provide different results depending on the model. On the other hand, models with the Densenet121 backbone obtain overall better results than models using VGG-16 for all individual loss functions. Using U-Net (concatenation) or LinkNet (add) lead to similar results for the individual loss functions.

Figures 2 and 3 show the boxplots for the accuracy and Jaccard index for the four models. Boxplots for the rest of metrics are provided as Supplementary Materials (Figures S1–S5). While the accuracy is more similar for all loss functions and models, the Jaccard index presents a greater variability and differences among them.

The correlation matrix presents high values ranging from 0.659 to 0.898 outside the main diagonal of the matrix for the four analyzed models. Therefore, correlation between the variables under study can be considered high and significant (p-value ≤ 0.001 in all cases). Furthermore, the value of the determinant of the correlation matrix was 0 for each model, so there is a linear relationship between the variables under study. Next, for the Bartlett sphericity test value, we obtain a χ^2 greater than 10,000 which indicates that there are high values outside the diagonal of the correlation matrix with significance less than 0.001, for all models. In the KMO sample fit, the value ranged from 0.937 to 0.941 for the four models, which corroborates the results of the previous statistical parameters. Therefore, we can conclude that the PCA can be applied to the variables under study in the four models.

The PCA analysis identified only one independent component for each considered model (Table 5).

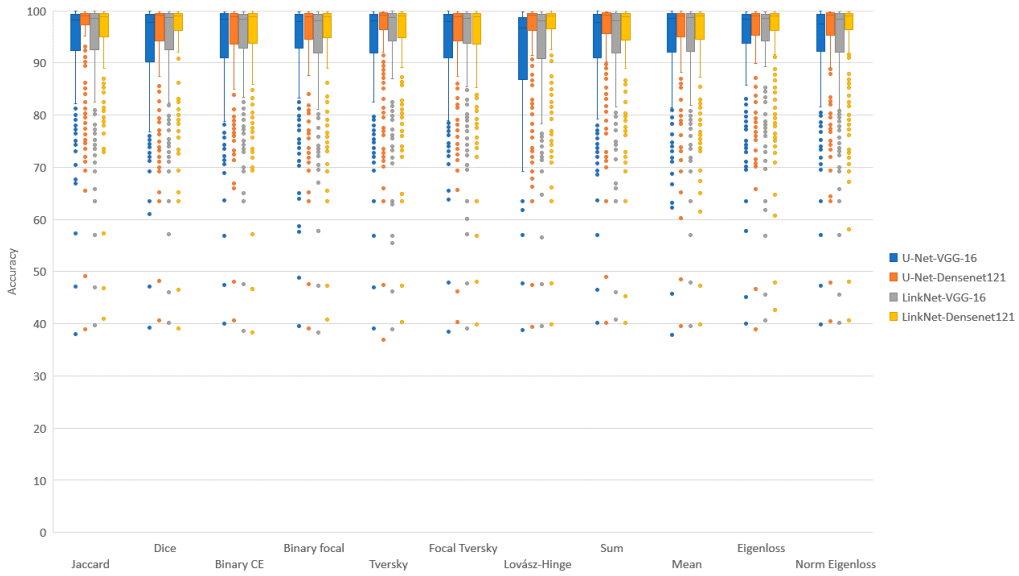


Figure 3. Boxplot for accuracy for the different models and loss functions analyzed.

Table 5. PCA results. Com: Communality; α_{1n}

	U-Net-VGG-16		**U-Net-Densenet121**		**LinkNet-VGG-16**		**LinkNet-Densenet121**	
Eigenvalue	5.803		6.076		5.913		5.998	
% Variance	82.91		86.80		84.47		85.69	
Loss function	**Com.**	α_{1n}	**Com.**	α_{1n}	**Com.**	α_{1n}	**Com.**	α_{1n}
Jaccard	0.875	0.935	0.819	0.905	0.843	0.918	0.819	0.905
Dice	0.845	0.919	0.894	0.946	0.864	0.930	0.878	0.937
Binary cross entropy	0.835	0.914	0.785	0.886	0.866	0.930	0.842	0.918
Binary focal	0.826	0.909	0.892	0.944	0.805	0.897	0.861	0.928
Tversky	0.881	0.939	0.906	0.952	0.842	0.917	0.881	0.939
Focal Tversky	0.884	0.940	0.869	0.932	0.888	0.942	0.878	0.937
Lovász-Hinge	0.657	0.811	0.911	0.954	0.805	0.897	0.840	0.916

Based on these results, the eigenloss was calculated using the coefficients of the first principal component, α_{1n}, from Table 5 in (Equation (17)) for each considered model. Results for the test set of the models trained with its corresponding eigenloss are provided in the penultimate row of Tables 1–4, as well as in the penultimate boxes of Figures 3 and 4. For comparison, we also analyzed other linear combinations:

- Sum: where all coefficients are equal to 1.
- Mean: where all coefficients are equal to 1/7.
- Normalized eigenloss: where coefficients α_{1n} are normalized using the formula.

$$\alpha'_{1n} = \frac{\alpha_{1n}}{\sum_n \alpha_{1n}} \tag{18}$$

with $n \in \{Jaccard,\ Dice,\ BCE, BF, Tversky, FocalTversky,\ Lovasz - Hinge\}$.

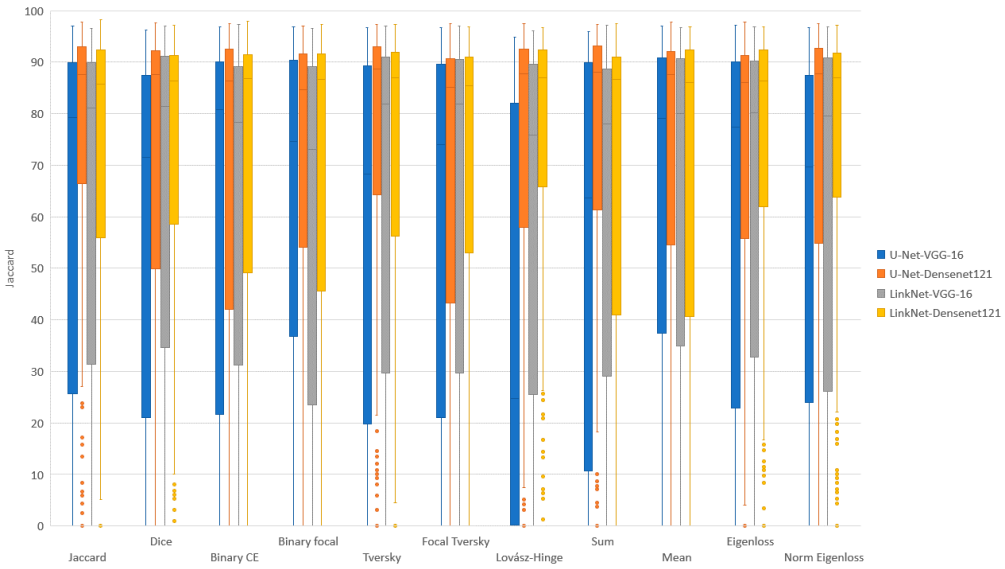


Figure 4. Boxplot for the Jaccard index for the different models and loss functions analyzed.

The proposed eigenloss improves the general performance of individual loss functions in all models, as it provides closer values to 100 in the seven metrics at the same time, although the maximum value might be for any of the individual losses. It is interesting to remark that the F2-score of the combined loss function improves, therefore providing a better balance between precision and recall than the individual loss functions. Our proposed combination also minimizes variability in the Jaccard index.

In comparison to other linear combinations, the eigenloss obtains better results than the sum and the mean when LinkNet architecture is considered, regardless of the backbone. On the other hand, for the U-Net, either the sum or the mean performs better.

3.2. Training and Testing Analysis

Table 6 summarizes the best epoch and the corresponding validation loss at which weights are selected for prediction as indicated in Section 2.2.

Models with VGG-16 converge slower, the lowest epoch is 172 and 211 for the U-Net-VGG-16 and LinkNet-VGG-16, respectively. On the other hand, LinkNet-Densenet121 converges faster than U-Net-Densenet121 for most losses. In general, all loss values in the best epochs are quite similar, ranging from 0.12 to 0.20. These findings are also observable in Figures S6–S9 included as Supplementary Materials. They show the evolution of the losses during training for the training and validation sets. Peaks are produced when the learning rate is restored to its initial value.

Although there are no great differences in the initial convergence within each model, largest differences are obtained in LinkNet-VGG-16, where Lovász-Hinge is the slowest one and the binary cross entropy the fastest one among the individual loss functions (Figure 5). In all cases, the general trend of the eigenloss is to run within the maximum and minimum values set by the individual loss functions.

Table 6. Best epoch (# Epoch) and value of the validation loss (Val Loss) for each model and loss function.

	U-Net-VGG-16		U-Net-Densenet121		LinkNet-VGG-16		LinkNet-Densenet121	
Loss Function	**# Epoch**	**Val Loss**	**# Epoch**	**Val Loss**	**# Epoch**	**Val Loss**	**# Epoch**	**Val Loss**
Jaccard	292	0.17	71	0.16	292	0.18	17	0.15
Dice	216	0.19	71	0.12	211	0.17	71	0.16
Binary cross entropy	292	0.16	71	0.15	292	0.15	211	0.14
Binary focal	198	0.20	153	0.19	292	0.15	17	0.16
Tversky	292	0.16	71	0.16	292	0.18	47	0.18
Focal Tversky	216	0.20	71	0.14	292	0.16	211	0.17
Lovász-Hinge	184	0.19	71	0.15	216	0.16	17	0.16
Sum	259	0.22	71	0.15	292	0.18	128	0.15
Mean	172	0.16	71	0.14	211	0.12	17	0.14
Eigenloss	211	0.19	114	0.17	211	0.14	17	0.16
Normalized eigenloss	211	0.20	71	0.13	292	0.16	17	0.18

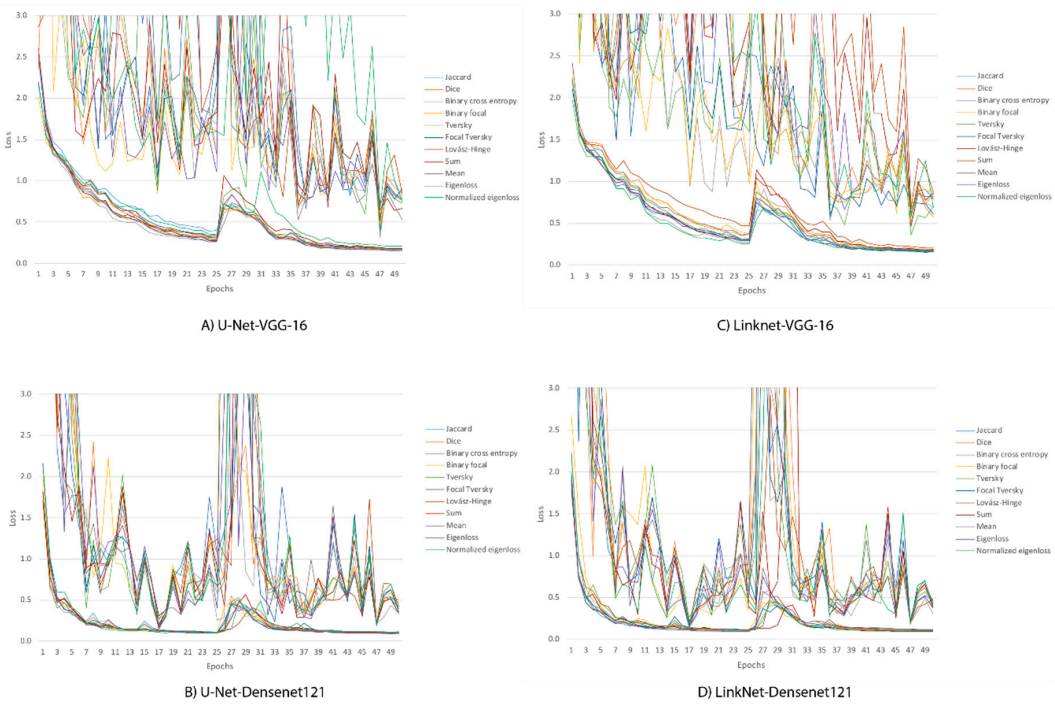


Figure 5. Initial convergence. (**A**) U-Net-VGG-16; (**B**) U-Net-Densenet121; (**C**) LinkNet-VGG-16; (**D**) LinkNet-Densenet121. Dark colors correspond to the training loss, while light colors correspond to the validation loss.

Figures S10–S13 provided in the Supplementary Materials show the evolution of the loss, accuracy and Jaccard index during training for the training and validation sets for each model and each loss function. In all cases, accuracy values for train and validation sets remain very similar from the first epochs, achieving values very close to 100. This is due to fact that the background class is considered in the accuracy, so even when the polyp is not detected at all, accuracy is still high thanks to the negative class. Therefore, accuracy is not the most suitable metric for polyp segmentation. On the other hand, Jaccard index values in training and validation set vary along the whole training for all models and loss functions. Moreover, it shows greater fluctuations, showing a more erratic behavior when VGG-16 is used as backbone. As it might be expected, Jaccard and Dice losses provide a more stable value for the Jaccard index during training for the training and validation set, as the metric is either directly or

indirectly included in the loss function for optimization. The proposed eigenloss remains acceptably stable during training, reaching values around 60 in the best epoch.

4. Discussion

Computer-aided detection systems have the potential to improve colonoscopy [6]. In this way, polyps, considered early precursors of CRC, can be detected, diagnosed and treated. In this work, we have faced the problem of polyp detection by the application of segmentation models based on deep learning. Specifically, we have analyzed the influence of seven loss functions on four models for polyp segmentation and determined whether a PCA-based decomposition allows for the defining of a non-redundant primal loss function that can outperform them and different linear combinations.

Through the application of PCA, the eigenloss was defined as a linear combination of seven individual loss functions. It does improve overall performance over the individual loss functions for the four considered models. On the other hand, and only for the LinkNet architecture, the eigenloss performs better than the sum and the mean. Regarding the backbones, the eigenlos performs better with the Densenet121 than VGG-16, in all cases. This might be due to the fact that VGG-16 presents a more linear architecture that cannot extract enough details for segmenting complex structures such as the polyp, due to the high level of visual similarity between healthy mucosa and lesion. Furthermore, it is also worth mentioning that the eigenloss provides the most balanced results in terms of precision and recall, therefore balancing false positives and false negatives.

Based on the experimental results of the individual loss functions, region-based loss functions provide better results for polyp segmentation. While highest values remain in the individual loss functions in many cases, the eigenloss improves the overall performance providing a more balanced result among all metrics for each model. Therefore, one benefit of using the eigenloss is to include the properties of both region and distribution-based loss functions in one single loss function, which might be the reason for this balanced behavior, which arises from the PCA capability for projecting the existing losses into a statistically independent eigenloss. This way, there is no need for choosing one over the other, except in the case that one loss function behaves outstandingly well in all metrics. Due to the fact that PCA provides a compression with information loss, the use of the eigenloss in this case might hinder the performance if the individual loss function is used.

Our work can be straightforwardly compared to other works that also used the CVC-EndoSceneStill dataset (Table 7), although regretfully, the rest of the authors do not provide all of the metrics included in this work. Our proposed eigenloss leads to better results in all but one of the five comparable metrics. Nevertheless, we acknowledge that properly ranking the methods is challenging and particular guidelines would help stablish a common framework for comparison [76].

Table 7. Comparison of the eigenloss of the four models with other works that also use CVC-EndoSceneStill as the dataset. Best values are indicated in bold.

Work	Accuracy	Precision	Recall	Specificity	F2-Score	Jaccard	Dice
U-Net-VGG-16	93.00 ± 10.91	80.31 ± 31.35	65.07 ± 37.70	99.00 ± 2.25	65.45 ± 36.70	59.66 ± 35.32	67.04 ± 35.39
U-Net-Densenet121	94.43 ± 9.78	**89.83 ± 21.63**	77.39 ± 33.28	99.06 ± 2.18	76.41 ± 31.97	69.35 ± 31.55	76.33 ± 30.39
LinkNet-VGG-16	93.50 ± 10.60	83.31 ± 28.49	69.52 ± 36.50	**99.28 ± 1.76**	69.51 ± 35.31	63.24 ± 33.86	70.61 ± 33.68
LinkNet-Densenet121	94.59 ± 9.61	88.65 ± 18.25	**80.23 ± 30.78**	99.05 ± 1.92	**79.18 ± 29.26**	**71.84 ± 29.25**	**79.07 ± 27.07**
Vázquez et al. [39]	**96.77**	-	-	-	-	56.07	-
Wichakam et al. [37]	-	88.84	78.14	-	-	69.36	78.61
Wickstrøm et al. [41]	94.90	-	-	-	-	58.70	-

Analyzing the full training, accuracy values for train and validation sets remains very similar from the first epochs, achieving values very close to 100 for all loss functions. Therefore, accuracy is not the most suitable metric for polyp segmentation mainly because of the imbalanced problem. In this situation, Taha et al. [77] recommend the use of distance-based metrics, such as the Hausdorff distance, average Hausdorff distance or the Mahalanobis distance.

This work also has limitations to be acknowledged. Only the CVC-EndoSceneStill dataset was used. This publicly available dataset has a limited size, thus it might be possible that other datasets behave differently, depending on the type and shape of polyps included in their images. Since the imbalance level of the dataset have a great influence on metrics, a different dataset where polyps are bigger or smaller might lead to a different combination. On the other hand, only seven metrics based on the confusion matrix were used for the PCA analysis. The inclusion of the previously mentioned distance-based metrics could also yield to different PCA results and therefore a different eigenloss.

Therefore, future works should firstly consider the use of different datasets for comparison as well as images of normal tissue (frames containing no polyps at all). Another future study prospect relies on the need for studying the influence of polyp size in the performance of the segmentation models and loss functions. We also consider the inclusion of more complex loss functions to the study, such as the rectified cross-entropy [78], random drop loss [79], the integrated Hausdorff-Sine loss [80] or the enhanced softmax loss [81], which were not considered in this study mainly due to time and resources constraints. Most of them are proposed for very specific cases, so it should be tested whether they are also suitable for polyp segmentation. Regarding the selected models, performance of other U-net based approaches could be further analyzed. These approaches have been proven to be highly effective for different biomedical imaging applications [26,82,83]. In this comparison, it might be also interesting to take into account prominent segmentation networks, such as DeepLab [84] or CCNet [85], which the performs well in natural images, in order to study their suitability for polyp segmentation. Lastly, it would also be interesting to establish a detailed analysis of transferability of PCA-based losses, in order to further evaluate whether coefficients for the eigenloss can be reused, analyzing whether the level of generalization is enough, or whether a new PCA analysis is required for each particular model.

5. Conclusions

In this work, we propose the use of PCA as a method to extract the coefficients of a linear combination of seven individual loss functions. We have considered two different encoder-decoder architectures with two different backbones. LinkNet-Densenet121 provides the best performance out of the four models. Regardless of the backbone, if LinkNet is used, the eigenloss results in more stable performance (all metrics have an acceptable value) in comparison to other linear combinations, such as the sum or the mean. Convergence of models using VGG-16 as the backbone is slower than those using Densener121. Further analysis is required to determine the transferability of PCA-based losses.

Supplementary Materials: The following are available online at http://www.mdpi.com/2227-7390/8/8/1316/s1, Table S1: U-Net-VGG-16 architecture, Table S2: U-Net-Densenet121 architecture, Table S3: LinkNet-VGG-16 architecture, Table S4: LinkNet-Densenet121 architecture. Figure S1: Boxplot for precision for the different loss functions and models analyzed, Figure S2: Boxplot for recall for the different loss functions and models analyzed, Figure S3: Boxplot for specificity for the different loss functions and models analyzed, Figure S4: Boxplot for F2-score for the different loss functions and models analyzed, Figure S5: Boxplot for Dice index for the different loss functions and models analyzed, Figure S6: Train and validation loss rates for U-Net-VGG-16 for all loss functions, Figure S7: Train and validation loss rates for U-Net-Densenet121 for all loss functions, Figure S8: Train and validation loss rates for LinkNet-VGG-16 for all loss functions, Figure S9: Train and validation loss rates for LinkNet-Densenet121 for all loss functions, Figure S10: Loss rate, accuracy and Jaccard index during training for the train and validation sets for the U-Net-VGG-16, Figure S11: Loss rate, accuracy and Jaccard index during training for the train and validation sets for the U-Net-Densenet121, Figure S12: Loss rate, accuracy and Jaccard index during training for the train and validation sets for the LinkNet-VGG-16, Figure S13: Loss rate, accuracy and Jaccard index during training for the train and validation sets for the LinkNet-Densenet121.

Author Contributions: Conceptualization, L.F.S.-P., A.P. and J.B.P.; methodology, L.F.S.-P., A.P., J.A.A.-B. and J.B.P.; software, L.F.S.-P. and A.P.; validation, L.F.S.-P., A.P. and J.B.P.; formal analysis, L.F.S.-P. and J.A.A.-B.; investigation, L.F.S.-P.; resources, F.M.S.-M. and J.B.P.; data curation, L.F.S.-P.; writing—original draft preparation, L.F.S.-P., J.F.O.-M. and J.B.P.; writing—review and editing, L.F.S.-P., J.A.A.-B., J.F.O.-M., A.P. and J.B.P.; visualization, L.F.S.-P.; supervision, A.P., F.M.S.-M. and J.B.P.; project administration, J.F.O.-M., F.M.S.-M. and J.B.P.; funding acquisition, A.P., F.M.S.-M. and J.B.P. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partially supported by PICCOLO project. This project has received funding from the European Union's Horizon2020 research and innovation programme under grant agreement No 732111. The sole

responsibility of this publication lies with the author. The European Union is not responsible for any use that may be made of the information contained therein.

Acknowledgments: The authors would like to thank Juan Carlos Gómez-Blanco for his help with the images of this work.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. International Agency for Research on Cancer. *World Cancer Report 2020*; Wild, C.P., Weiderpass, E., Stewart, B.W., Eds.; International Agency for Research on Cancer: Lyon, France, 2020; ISBN 9789283204299.
2. International Agency for Research on Cancer. *Colorectal Cancer Factsheet*; International Agency for Research on Cancer: Lyon, France, 2018.
3. Siegel, R.L.; Miller, K.D.; Jemal, A. Cancer statistics, 2020. *CA Cancer J. Clin.* **2020**, *70*, 7–30. [CrossRef]
4. Carioli, G.; Bertuccio, P.; Boffetta, P.; Levi, F.; La Vecchia, C.; Negri, E.; Malvezzi, M. European cancer mortality predictions for the year 2020 with a focus on prostate cancer. *Ann. Oncol.* **2020**, *31*, 650–658. [CrossRef] [PubMed]
5. Digestive Cancers Europe. *Colorectal Screening in Europe Saving Lives and Saving Money*; Digestive Cancers Europe: Brussles, Belgium, 2019; Available online: https://www.digestivecancers.eu/wp-content/uploads/2020/02/466-Document-DiCEWhitePaper2019.pdf (accessed on 1 August 2020).
6. Byrne, M.F.; Shahidi, N.; Rex, D.K. Will Computer-Aided Detection and Diagnosis Revolutionize Colonoscopy? *Gastroenterology* **2017**, *153*, 1460–1464.e1. [CrossRef] [PubMed]
7. Chen, Y.-W.; Jain, L.C. (Eds.) *Deep Learning in Healthcare*; Intelligent Systems Reference Library; Springer International Publishing: Cham, Switzerland, 2020; Volume 171, ISBN 978-3-030-32605-0.
8. Tajbakhsh, N.; Jeyaseelan, L.; Li, Q.; Chiang, J.N.; Wu, Z.; Ding, X. Embracing Imperfect Datasets: A Review of Deep Learning Solutions for Medical Image Segmentation. *Med. Image Anal.* **2020**, *63*, 101693. [CrossRef]
9. Nagendran, M.; Chen, Y.; Lovejoy, C.A.; Gordon, A.C.; Komorowski, M.; Harvey, H.; Topol, E.J.; Ioannidis, J.P.A.; Collins, G.S.; Maruthappu, M. Artificial intelligence versus clinicians: Systematic review of design, reporting standards, and claims of deep learning studies in medical imaging. *BMJ* **2020**, *368*, 1–12. [CrossRef] [PubMed]
10. Liu, X.; Faes, L.; Kale, A.U.; Wagner, S.K.; Fu, D.J.; Bruynseels, A.; Mahendiran, T.; Moraes, G.; Shamdas, M.; Kern, C.; et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: A systematic review and meta-analysis. *Lancet Digit. Health* **2019**, *1*, e271–e297. [CrossRef]
11. E Gonçalves, W.G.; dos Santos, M.H.D.P.; Lobato, F.M.F.; Ribeiro-dos-Santos, Â.; de Araújo, G.S. De Deep learning in gastric tissue diseases: A systematic review. *BMJ Open Gastroenterol.* **2020**, *7*, e000371. [CrossRef]
12. Soffer, S.; Klang, E.; Shimon, O.; Nachmias, N.; Eliakim, R.; Ben-Horin, S.; Kopylov, U.; Barash, Y. Deep learning for wireless capsule endoscopy: A systematic review and meta-analysis. *Gastrointest. Endosc.* **2020**. [CrossRef]
13. Iwahori, Y.; Hagi, H.; Usami, H.; Woodham, R.J.; Wang, A.; Bhuyan, M.K.; Kasugai, K. Automatic polyp detection from endoscope image using likelihood map based on edge information. In *ICPRAM 2017—Proceedings 6th International Conference Pattern Recognit, Porto, Portugal, 24-26 February 2017*; SciTePress: Setubal, Portugal, 26 February; pp. 402–409. [CrossRef]
14. Iakovidis, D.K.; Maroulis, D.E.; Karkanis, S.A. An intelligent system for automatic detection of gastrointestinal adenomas in video endoscopy. *Comput. Biol. Med.* **2006**, *36*, 1084–1103. [CrossRef]
15. Ameling, S.; Wirth, S.; Paulus, D.; Lacey, G.; Vilarino, F. Texture-based polyp detection in colonoscopy. In Proceedings of the Bildverarbeitung für die Medizin 2009, Heidelberg, Germany, 22–25 March 2009; pp. 346–350.
16. Bernal, J.; Sánchez, F.J.; Fernández-Esparrach, G.; Gil, D.; Rodríguez de Miguel, C.; Vilariño, F. WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Comput. Med. Imaging Graph.* **2015**, *43*, 99–111. [CrossRef]
17. Bernal, J.; Sánchez, F.J.; Vilariño, F. Towards automatic polyp detection with a polyp appearance model. *Pattern Recognit.* **2012**, *45*, 3166–3182. [CrossRef]

18. Minaee, S.; Wang, Y. An ADMM Approach to Masked Signal Decomposition Using Subspace Representation. *IEEE Trans. Image Process.* **2019**, *28*, 3192–3204. [CrossRef] [PubMed]
19. Li, J.; Bioucas-Dias, J.M.; Plaza, A. Semisupervised hyperspectral image segmentation using multinomial logistic regression with active learning. *IEEE Trans. Geosci. Remote Sens.* **2010**, *48*, 4085–4098. [CrossRef]
20. Lui, T.K.; Hui, C.K.; Tsui, V.W.; Cheung, K.S.; Ko, M.K.; aCC Foo, D.; Mak, L.Y.; Yeung, C.K.; Lui, T.H.; Wong, S.Y.; et al. New insights on missed colonic lesions during colonoscopy through artificial intelligence–assisted real-time detection (with video). *Gastrointest. Endosc.* **2020**. [CrossRef] [PubMed]
21. Bernal, J.; Tajbakhsh, N.; Sánchez, F.J.; Matuszewski, B.J.; Chen, H.; Yu, L.; Angermann, Q.; Romain, O.; Rustad, B.; Balasingham, I.; et al. Comparative Validation of Polyp Detection Methods in Video Colonoscopy: Results from the MICCAI 2015 Endoscopic Vision Challenge. *IEEE Trans. Med. Imaging* **2017**, *36*, 1231–1249. [CrossRef] [PubMed]
22. Sánchez-Peralta, L.F.; Bote-Curiel, L.; Picon, A.; Sánchez-Margallo, F.M.; Pagador, J.B. Deep learning to find colorectal polyps in colonoscopy: A systematic literature review. *Artif. Intell. Med.* **2020**, in press. [CrossRef]
23. Minaee, S.; Boykov, Y.; Porikli, F.; Plaza, A.; Kehtarnavaz, N.; Terzopoulos, D. Image Segmentation Using Deep Learning: A Survey. *arXiv* **2020**, arXiv:2001.05566.
24. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*; Lecture Notes in Computer Science, vol 9351; Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F., Eds.; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241. ISBN 9783319245737.
25. Milletari, F.; Navab, N.; Ahmadi, S. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. In Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV), Stanford, CA, USA, 25–28 October 2016; pp. 565–571. [CrossRef]
26. Zhou, Z.; Siddiquee, M.M.R.; Tajbakhsh, N.; Liang, J. UNet++: A Nested U-Net Architecture for Medical Image Segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*; DLMIA 2018, ML-CDS 2018; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2018; Volume 11045, ISBN 1424400430.
27. Zahangir Alom, M.; Yakopcic, C.; Taha, T.M.; Asari, V.K. Nuclei Segmentation with Recurrent Residual Convolutional Neural Networks based U-Net (R2U-Net). In Proceedings of the NAECON 2018-IEEE National Aerospace and Electronics Conference, Dayton, OH, USA, 23–26 July 2018; pp. 228–233. [CrossRef]
28. Isensee, F.; Petersen, J.; Klein, A.; Zimmerer, D.; Jaeger, P.F.; Kohl, S.; Wasserthal, J.; Koehler, G.; Norajitra, T.; Wirkert, S.; et al. nnU-Net: Self-adapting Framework for U-Net-Based Medical Image Segmentation. *Inform. Aktuell* **2019**, 22. [CrossRef]
29. Chaurasia, A.; Culurciello, E. LinkNet: Exploiting encoder representations for efficient semantic segmentation. In Proceedings of the 2017 IEEE Visual Communications and Image Processing (VCIP), St. Petersburg, FL, USA, 10–13 December 2018. [CrossRef]
30. Rezaei, S.; Emami, A.; Zarrabi, H.; Rafiei, S.; Najarian, K.; Karimi, N.; Samavi, S.; Reza Soroushmehr, S.M. Gland Segmentation in Histopathology Images Using Deep Networks and Handcrafted Features. In Proceedings of the 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Berlin, Germany, 23–27 July 2019; pp. 1031–1034. [CrossRef]
31. Lee, J.Y.; Islam, M.; Woh, J.R.; Washeem, T.S.M.; Ngoh, L.Y.C.; Wong, W.K.; Ren, H. Ultrasound needle segmentation and trajectory prediction using excitation network. *Int. J. Comput. Assist. Radiol. Surg.* **2020**, *15*, 437–443. [CrossRef]
32. Bagheri, M.; Mohrekesh, M.; Tehrani, M.; Najarian, K.; Karimi, N.; Samavi, S.; Reza Soroushmehr, S.M. Deep Neural Network based Polyp Segmentation in Colonoscopy Images using a Combination of Color Spaces. In Proceedings of the 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Berlin, Germany, 23–27 July 2019; pp. 6742–6745. [CrossRef]
33. Garcia-Pedrero, A.; García-Cervigón, A.I.; Olano, J.M.; García-Hidalgo, M.; Lillo-Saavedra, M.; Gonzalo-Martín, C.; Caetano, C.; Calderón-Ramírez, S. Convolutional neural networks for segmenting xylem vessels in stained cross-sectional images. *Neural Comput. Appl.* **2019**, *6*. [CrossRef]
34. Shvets, A.A.; Rakhlin, A.; Kalinin, A.A.; Iglovikov, V.I. Automatic Instrument Segmentation in Robot-Assisted Surgery using Deep Learning. In Proceedings of the 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), Orlando, FL, USA, 17–20 December 2019; pp. 624–628. [CrossRef]

35. Kholiavchenko, M.; Sirazitdinov, I.; Kubrak, K.; Badrutdinova, R.; Kuleev, R.; Yuan, Y.; Vrtovec, T.; Ibragimov, B. Contour-aware multi-label chest X-ray organ segmentation. *Int. J. Comput. Assist. Radiol. Surg.* **2020**, *15*, 425–436. [CrossRef] [PubMed]
36. Singh, J.; Tripathy, A.; Garg, P.; Kumar, A. Lung tuberculosis detection using anti-aliased convolutional networks. *Procedia Comput. Sci.* **2020**, *173*, 281–290. [CrossRef]
37. Wichakam, I.; Panboonyuen, T.; Udomcharoenchaikit, C. Real-Time Polyps Segmentation for Colonoscopy Video Frames Using Compressed Fully Convolutional Network. In *International Conference on Multimedia Modeling*; Lecture Notes in Computer Science, vol 10704; Springer: Cham, Switzerlan, 2018; pp. 393–404. ISBN 978-3-319-51813-8.
38. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2015**, arXiv:1409.1556. [CrossRef]
39. Vázquez, D.; Bernal, J.; Sánchez, F.J.; Fernández-Esparrach, G.; López, A.M.; Romero, A.; Drozdzal, M.; Courville, A. A Benchmark for Endoluminal Scene Segmentation of Colonoscopy Images. *J. Healthc. Eng.* **2017**. [CrossRef]
40. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
41. Wickstrøm, K.; Kampffmeyer, M.; Jenssen, R. Uncertainty modeling and interpretability in convolutional neural networks for polyp segmentation. In Proceedings of the 2018 IEEE International Workshop on Machine Learning for Signal Processing, Aalborg, Denmark, 17–20 September 2018.
42. Jolliffe, I.T. *Principal Components Analysis*; Springer: New York, NY, USA, 2002; ISBN 978-0-387-22440-4.
43. Ansari, K.; Krebs, A.; Benezeth, Y.; Marzani, F. Color Converting of Endoscopic Images Using Decomposition Theory and Principal Component Analysis. In Proceedings of the 9th International Conference on Computer Science, Engineering and Applications, Toronto, ON, Canada, 13–14 July 2019; pp. 151–159. [CrossRef]
44. Shao, X.; Zheng, W.; Huang, Z. Near-infrared autofluorescence spectroscopy for in vivo identification of hyperplastic and adenomatous polyps in the colon. *Biosens. Bioelectron.* **2011**, *30*, 118–122. [CrossRef]
45. Kim, Y.; Kim, H.G.; Hyeon, J.; Choi, H.J. Clinical opinions generation from general blood test results using deep neural network with principle component analysis and regularization. In Proceedings of the 2017 IEEE International Conference on Big Data and Smart Computing (BigComp), Jeju, Korea, 13–16 February 2017; pp. 386–389. [CrossRef]
46. Sánchez-González, A.; García-Zapirain, B.; Sierra-Sosa, D.; Elmaghraby, A. Automatized colon polyp segmentation via contour region analysis. *Comput. Biol. Med.* **2018**, *100*, 152–164. [CrossRef]
47. Cho, K.; Roh, J.H.; Kim, Y.; Cho, S. A Performance Comparison of Loss Functions. In Proceedings of the 2019 International Conference on Information and Communication Technology Convergence (ICTC), Jeju Island, Korea, 16–18 October 2019; pp. 1146–1151. [CrossRef]
48. Keren, G.; Sabato, S.; Schuller, B. Analysis of loss functions for fast single-class classification. *Knowl. Inf. Syst.* **2020**, *62*, 337–358. [CrossRef]
49. Ghodrati, V.; Shao, J.; Bydder, M.; Zhou, Z.; Yin, W.; Nguyen, K.L.; Yang, Y.; Hu, P. MR image reconstruction using deep learning: Evaluation of network structure and loss functions. *Quant. Imaging Med. Surg.* **2019**, *9*, 1516–1527. [CrossRef]
50. Kim, B.; Han, M.; Shim, H.; Baek, J. A performance comparison of convolutional neural network-based image denoising methods: The effect of loss functions on low-dose CT images. *Med. Phys.* **2019**, *46*, 3906–3923. [CrossRef]
51. Pathak, A.; Maheshwari, R. Comparative analysis of different loss functions for deep face recognition. In Proceedings of the 2019 2nd International Conference on Algorithms, Computing and Artificial Intelligence, Sanya, China, 20–22 December 2019; pp. 390–397. [CrossRef]
52. Sukhbaatar, S.; Bruna, J.; Paluri, M.; Bourdev, L.; Fergus, R. Training convolutional networks with noisy labels. *arXiv* **2014**, arXiv:1406.2080.
53. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollar, P. Focal Loss for Dense Object Detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2999–3007. [CrossRef]

54. Rahman, M.A.; Wang, Y. Optimizing intersection-over-union in deep neural networks for image segmentation. In *International Symposium on Visual Computing*; 10072 LNCS; Springer: Cham, Switzerland, 2016; pp. 234–244. [CrossRef]
55. Salehi, S.S.M.; Erdogmus, D.; Gholipour, A. Tversky loss function for image segmentation using 3D fully convolutional deep networks. In *International Workshop on Machine Learning in Medical Imaging*; 10541 LNCS; Springer: Cham, Switzerland, 2017; pp. 379–387. [CrossRef]
56. Berman, M.; Triki, A.R.; Blaschko, M.B. The Lovasz-Softmax Loss: A Tractable Surrogate for the Optimization of the Intersection-Over-Union Measure in Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 4413–4421. [CrossRef]
57. Abraham, N.; Khan, N.M. A novel focal tversky loss function with improved attention u-net for lesion segmentation. In Proceedings of the 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), Venice, Italy, 8–11 April 2019; pp. 683–687. [CrossRef]
58. Kervadec, H.; Bouchtiba, J.; Desrosiers, C.; Granger, É.; Dolz, J.; Ayed, I. Ben Boundary loss for highly unbalanced segmentation. In Proceedings of the International Conference on Medical Imaging with Deep Learning, London, UK, 8–10 July 2019; pp. 285–296.
59. Karimi, D.; Salcudean, S.E. Reducing the Hausdorff Distance in Medical Image Segmentation with Convolutional Neural Networks. *IEEE Trans. Med. Imaging* **2020**, *39*, 499–513. [CrossRef]
60. Asgari Taghanaki, S.; Abhishek, K.; Cohen, J.P.; Cohen-Adad, J.; Hamarneh, G. *Deep Semantic Segmentation of Natural and Medical Images: A Review*; Springer: Cham, The Netherlands, 2020; ISBN 0123456789.
61. Hajiabadi, H.; Babaiyan, V.; Zabihzadeh, D.; Hajiabadi, M. Combination of loss functions for robust breast cancer prediction. *Comput. Electr. Eng.* **2020**, *84*, 106624. [CrossRef]
62. Wang, B.; Lei, Y.; Tian, S.; Wang, T.; Liu, Y.; Patel, P.; Ashesh, B.; Mao, H.; Curran, W.J.; Liu, T.; et al. Deeply Supervised 3D FCN with Group Dilated Convolution for Automatic MRI Prostate Segmentation. *Med. Phys.* **2019**, *46*, 1707–1718. [CrossRef] [PubMed]
63. Oksuz, I.; Clough, J.; Ruijsink, B.; Puyol-Antón, E.; Gastao Cruz, K.; Prieto, C.; King, A.P.; Schnabel, J.A. High-quality segmentation of low quality cardiac MR images using k-space artefact correction. In Proceedings of the International Conference on Medical Imaging with Deep Learning, London, UK, 8–10 July 2019; Volume 102, pp. 380–389.
64. Mohammed, A.; Yildirim, S.; Farup, I.; Pedersen, M.; Hovde, Ø. Y-Net: A deep Convolutional Neural Network for Polyp Detection. *arXiv* **2018**, arXiv:1806.01907.
65. Huang, G.; Liu, Z.; van der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018. [CrossRef]
66. Stanford Vision Lab. Stanford University. Princeton University ImageNet. Available online: http://www.image-net.org/ (accessed on 10 February 2020).
67. Yakubovskiy, P. Segmentation Models. 2019. Available online: https://segmentation-models.readthedocs.io/en/latest/index.html (accessed on 1 August 2020).
68. Chollet, F. Keras. Available online: https://github.com/keras-team/keras (accessed on 10 February 2020).
69. Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G.S.; Davis, A.; Dean, J.; Devin, M.; et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *arXiv* **2016**, arXiv:1603.04467.
70. García-Gómez, J.M.; Tortajada, S. Definition of Loss Functions for Learning from Imbalanced Data to Minimize Evaluation metrics. In *Data Mining in Clinical Medicine*; Fernández-Llatas, C., García-Gómez, J.M., Eds.; Humana Press: New York, NY, USA, 2015; pp. 19–37.
71. Jaccard, P. The distribution of the flora in the alpine zone. *New Phytol.* **1912**, *11*, 37–50. [CrossRef]
72. Dice, L.R. Measures of the Amount of Ecologic Association Between Species. *Ecology* **1945**, *26*, 297–302. [CrossRef]
73. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016.
74. Tversky, A. Features of similarity. *Psychol. Rev.* **1977**, *84*, 327–352. [CrossRef]
75. Cliff, N. The Eigenvalues-Greater-Than-One Rule and the Reliability of Components. *Psychol. Bull.* **1988**, *103*, 276–279. [CrossRef]

76. Maier-Hein, L.; Eisenmann, M.; Reinke, A.; Onogur, S.; Stankovic, M.; Scholz, P.; Arbel, T.; Bogunovic, H.; Bradley, A.P.; Carass, A.; et al. Why rankings of biomedical image analysis competitions should be interpreted with care. *Nat. Commun.* **2018**, *9*, 5217. [CrossRef]
77. Taha, A.A.; Hanbury, A. Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool. *BMC Med. Imaging* **2015**, *15*. [CrossRef] [PubMed]
78. Qin, Z.; Zhang, Z.; Li, Y.; Guo, J. Making Deep Neural Networks Robust to Label Noise: Cross-Training with a Novel Loss Function. *IEEE Access* **2019**, *7*, 130893–130902. [CrossRef]
79. Guo, S.; Li, T.; Zhang, C.; Li, N.; Kang, H.; Wang, K. Random Drop Loss for Tiny Object Segmentation: Application to Lesion Segmentation in Fundus Images. In *International Conference on Artificial Neural Networks*; Springer: Cham, Switzerland, 2019; pp. 213–224.
80. Asaturyan, H.; Thomas, E.L.; Fitzpatrick, J.; Bell, J.D.; Villarini, B. Advancing Pancreas Segmentation in Multi-protocol MRI Volumes Using Hausdorff-Sine Loss Function. In *International Workshop on Machine Learning in Medical Imaging*; Springer: Cham, Switzerland, 2019; Volume 11861 LNCS, pp. 27–35.
81. Maharjan, S.; Alsadoon, A.; Prasad, P.W.C.; Al-Dalain, T.; Alsadoon, O.H. A novel enhanced softmax loss function for brain tumour detection using deep learning. *J. Neurosci. Methods* **2020**, *330*, 108520. [CrossRef] [PubMed]
82. Iglovikov, V.; Shvets, A. TernausNet: U-Net with VGG11 Encoder Pre-Trained on ImageNet for Image Segmentation. *arXiv* **2018**, arXiv:1801.05746.
83. Narayanan, B.N.; Hardie, R.C. A Computationally Efficient U-Net Architecture for Lung Segmentation in Chest Radiographs. In Proceedings of the 2019 IEEE National Aerospace and Electronics Conference (NAECON), Dayton, OH, USA, 15–19 July 2019; pp. 279–284. [CrossRef]
84. Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *arXiv* **2016**, arXiv:1606.00915. [CrossRef]
85. Huang, Z.; Wang, X.; Huang, L.; Huang, C.; Wei, Y.; Liu, W. CCNet: Criss-cross attention for semantic segmentation. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27 October–3 November 2019; pp. 603–612. [CrossRef]

© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).

8.5 PICCOLO white-light and narrow-band imaging colonoscopic dataset: a performance comparative of models and datasets

Dr. J. Blas Pagador Carrasco y Dr. Artzai Picón Ruiz, como directores de la tesis titulada "Mejoras en la detección, localización y segmentación automática de pólipos colorrectales basadas en técnicas de aprendizaje profundo", certificamos el factor de impacto y la categorización de la siguiente publicación, incluida en la tesis doctoral. Del mismo modo, se especifica la aportación del doctorando.

Autores: **Luisa F. Sánchez-Peralta**, J. Blas Pagador, Artzai Picón, Ángel José Calderón, Francisco Polo, Nagore Andraka, Roberto Bilbao, Ben Glover, Cristina L. Saratxaga, Francisco M. Sánchez-Margallo

Título: PICCOLO White-Light and Narrow-Band Imaging Colonoscopic Dataset: A Performance Comparative of Models and Datasets

Revista: Applied Sciences – Número especial: Deep Learning Applied to Image Processing

Otros datos: vol 10, nº 23, pág. 8501, año 2020

DOI: 10.3390/app10238501

Factor de impacto (2020): 2.679 (Q2 – Engineering, Multidisciplinary, Q3 – Chemistry, Multidisciplinary, Q2 – Physics, Applied, Q3 – Materials Science, Multidisciplinary)

Abstract: Colorectal cancer is one of the world leading death causes. Fortunately, an early diagnosis allows for effective treatment, increasing the survival rate. Deep learning techniques have shown their utility for increasing the adenoma detection rate at colonoscopy, but a dataset is usually required so the model can automatically learn features that characterize the polyps. In this work, we present the PICCOLO dataset, that comprises 3433 manually annotated images (2131 white-light images 1302 narrow-band images), originated from 76 lesions from 40 patients, which are distributed into training (2203), validation (897) and test (333) sets assuring patient independence between sets. Furthermore, clinical metadata are also provided for each lesion. Four different models, obtained by combining two backbones and two encoder–decoder architectures, are trained with the PICCOLO dataset and other two publicly available datasets for comparison. Results are provided for the test set of each dataset. Models trained with the PICCOLO dataset have a better generalization capacity, as they perform more uniformly along test sets of all datasets, rather than obtaining the best results for its own test set. This dataset is available at the website of the Basque Biobank, so it is expected that it will contribute to the further development of deep learning methods for polyp detection, localisation and classification, which would eventually result in a better and earlier diagnosis of colorectal cancer, hence improving patient outcomes.

Contribución del doctorando: Conceptualization, methodology, software, validation, formal analysis, investigation, data curation, writing—original draft preparation, writing—review and editing, visualization

Dr. D. José Blas Pagador Carrasco Dr. D. Artzai Picón Ruiz

© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).

applied sciences

MDPI

Article

PICCOLO White-Light and Narrow-Band Imaging Colonoscopic Dataset: A Performance Comparative of Models and Datasets

Luisa F. Sánchez-Peralta [1,*], J. Blas Pagador [1], Artzai Picón [2], Ángel José Calderón [3], Francisco Polo [3], Nagore Andraka [4], Roberto Bilbao [4], Ben Glover [5], Cristina L. Saratxaga [2] and Francisco M. Sánchez-Margallo [1]

1 Jesús Usón Minimally Invasive Surgery Centre, N-521, km 41.7, E-10071 Cáceres, Spain; jbpagador@ccmijesususon.com (J.B.P.); msanchez@ccmijesususon.com (F.M.S.-M.)
2 TECNALIA, Basque Research and Technology Alliance (BRTA), Parque Tecnológico de Bizkaia, C/Geldo. Edificio 700, E-48160 Derio-Bizkaia, Spain; artzai.picon@tecnalia.com (A.P.); Cristina.Lopez@tecnalia.com (C.L.S.)
3 Gastroenterology Department, Hospital Universitario Basurto, Avenida Montevideo, 18. E-48013 Bilbao, Spain; angeljose.calderongarcia@osakidetza.eus (Á.J.C.); francisco.poloortiz@osakidetza.eus (F.P.)
4 Basque Biobank, Basque Foundation for Health Innovation and Research-BIOEF, Ronda de Azkue 1, E-48902 Barakaldo, Spain; nagoreandraka@gmail.com (N.A.); bilbao@bioef.eus (R.B.)
5 Imperial College London, Exhibition Road, South Kensington, London SW7 2BU, UK; bglover@ic.ac.uk
* Correspondence: lfsanchez@ccmijesususon.com; Tel.: +34-927-18-10-32

Received: 30 October 2020; Accepted: 22 November 2020; Published: 28 November 2020

check for updates

Featured Application: This dataset can be used for supervised training of models for colorectal polyp detection, localisation, segmentation and classification.

Abstract: Colorectal cancer is one of the world leading death causes. Fortunately, an early diagnosis allows for effective treatment, increasing the survival rate. Deep learning techniques have shown their utility for increasing the adenoma detection rate at colonoscopy, but a dataset is usually required so the model can automatically learn features that characterize the polyps. In this work, we present the PICCOLO dataset, that comprises 3433 manually annotated images (2131 white-light images 1302 narrow-band images), originated from 76 lesions from 40 patients, which are distributed into training (2203), validation (897) and test (333) sets assuring patient independence between sets. Furthermore, clinical metadata are also provided for each lesion. Four different models, obtained by combining two backbones and two encoder–decoder architectures, are trained with the PICCOLO dataset and other two publicly available datasets for comparison. Results are provided for the test set of each dataset. Models trained with the PICCOLO dataset have a better generalization capacity, as they perform more uniformly along test sets of all datasets, rather than obtaining the best results for its own test set. This dataset is available at the website of the Basque Biobank, so it is expected that it will contribute to the further development of deep learning methods for polyp detection, localisation and classification, which would eventually result in a better and earlier diagnosis of colorectal cancer, hence improving patient outcomes.

Keywords: deep learning; colorectal cancer; public dataset; clinical metadata; colonoscopy; binary masks; polyps; detection; localisation; segmentation

1. Introduction

Colorectal Cancer (CRC) represents a 10% of overall new cases and presents higher incidence rate in developed countries [1] and could be considered a "lifestyle" disease associated with a diet high in calories and animal fat, and sedentarism [2]. In the United States, it has increased from over 132,000 estimated new cases and nearly 50,000 estimated deaths in 2015 [3] to 147,000 estimated new cases and 53,200 estimated deaths in 2020, being the third most commonly diagnosed cancer type [4]. Despite this, CRC detection increases the 5-year survival rate from 18% to 88.5% if diagnosed at an early stage. Furthermore, screening programs allow for detection before the appearance of symptoms so up to 22% of symptomatic cases could be treated earlier [5]. Colonoscopy is the gold standard procedure for detection and treatment of colorectal lesions, and its efficacy is related to the Adenoma Detection Rate (ADR) of the endoscopist, defined as the percentage of colonoscopies with at least one adenoma identified [6]. It is shown that higher ADR is associated with lower interval CRC rates [7], and that flat/sessile and small lesions are frequently more missed than pedunculated/sub-pedunculated and large ones [8,9]. Therefore, different approaches might be followed to improve ADR and reduce the number of missed lesions. During colonoscopy, these approaches include endoscope caps, positional manoeuvres, as well as the use of imaging modalities such as narrow-band imaging (NBI) which emphasize the capillary pattern and mucosa surface which emphasize the capillary pattern and mucosa surface [10]. It is clear that further development of Computer Assisted Diagnosis (CAD) systems is justified thanks to their potential to eventually improve the patient outcome [11].

In recent years, artificial intelligence (AI) and deep learning (DL) [12] have significantly contributed to the field of medical imaging analysis [13–15] and the use of AI in colonoscopy has showed promising results to increase ADR, although it should be further evaluated in more randomized controlled trials [16]. In the same trend, DL has also boosted the appearance of methods for polyp detection, localisation and segmentation, where end-to-end methods based on convolutional neural networks accompanied by data augmentation strategies are frequently used [17]. Nevertheless, methods for polyp detection are currently more advanced than methods for polyp classification, which might be due to a lack of available datasets for this task [18]. Lastly, it is important to mention that the availability of high-quality endoscopic images and the increased understanding of the technology by the endoscopists are two important factor for the further development of DL for endoscopy [19].

All deep learning approaches rely on a dataset based from which features can be learnt. If the training method is supervised, then the dataset should be labelled. Alternatively, semi-supervised training makes the most of labelled and unlabelled datasets. In the case of medical imaging datasets [20], data are hard to retrieve from healthcare systems, so it is necessary to obtain ethical approval and differently to labelling natural images, manually annotation of medical images is a cumbersome, time-consuming process that requires expert knowledge [21]. Especially when segmentation is the target, two main limitations are also usually recognized: scarce annotations and weak annotations [22]. Therefore, labelled datasets are usually smaller than non-labelled datasets, but in both cases, they are difficult to obtain, and, in many cases, they are proprietary datasets not publicly available for the research community.

There are currently some publicly available datasets of colonoscopy images that can be used for polyp detection, localisation and segmentation [17,18]. A general trend is that the more precise the annotation is, the smaller the dataset size is. In the matter at hand, annotation of manual precise binary masks is more time demanding than just labelling frames. The size of the publicly available datasets might range from hundreds of frames when a precise manually segmented binary mask is provided, to thousands of video frames if an approximated binary mask is created. The dataset CVC-EndosceneStill [23] provides manually segmented binary masks for polyp, background, lumen and specular lights classes. In all, 912 images are available. It also splits the dataset into training, validation and test sets and indicates the recommended metrics, so providing a common dataset which facilitates the proper comparison of methods. Kvasir-SEG [24] includes 1000 polyp images for which a polygonal binary mask and a bounding box are provided. On the other hand, CVC-VideoClinicDB [21,25] provides

an elliptical approximation for over 30,000 frames. Nevertheless, these datasets lack clinical metadata, such as the polyp size or histological diagnosis and only provide the Paris classification [26,27] in the best of the cases. Just recently, HyperKvasir [20] has been released. It includes labelled data for supervised training, but also provides unlabelled data. Regarding polyps, besides the images and binary masks of the Kvasir-SEG dataset, it also includes 1028 images and 73 videos labelled with the "polyp" class and 99,417 unlabelled images. All these datasets include white light (WL) images, but not NBI images.

For purposes of polyp classification, only one publicly available dataset is available [28]. This dataset includes 76 colonoscopy videos, using both WL and NBI imaging, of 15 serrated adenomas, 21 hyperplastic lesions and 40 adenomas.

The objective of this paper is to present the PICCOLO dataset with its associated clinical metadata and compare the performance results of different deep learning models trained with it and other publicly available datasets, as well as analyse the influence of the polyp morphology in the results. This paper is organized as follows: Section 2 details the acquisition and annotation protocols to obtain the PICCOLO dataset, as well as the public datasets and networks used in this study. In Section 3, we present and discuss the results of the experiments. Lastly, conclusions of this work are included in Section 4.

2. Materials and Methods

2.1. PICCOLO Dataset

The PICCOLO dataset contains several annotated frames from colonoscopy videos together with clinical metadata. In this dataset, both WL and NBI imaging technologies are included. The following subsections describe the acquisition and annotation protocols to generate the dataset.

2.1.1. Acquisition Protocol

An acquisition protocol was followed to obtain relevant information at Hospital Universitario Basurto (Bilbao, Spain):

1. Patients included in the colon cancer screening and surveillance program were informed about the study and asked for permission to collect images/videos obtained during routine colonoscopy with the associated information via informed consent and patient information sheet. If the patient gave permission, the rest of the protocol was followed.
2. If a suspicious lesion was found during the procedure, it was resected using the most appropriate method and sent to the Department of Pathological Anatomy for diagnosis.
3. Images/videos with the associated clinical information were anonymized.
4. Videos were analysed and edited/processed by the gastroenterologists who performed the colonoscopy. The video of the full procedure was divided into fragments, indicating if the tissue was healthy or pathological and the region on the colon where it was found. Further details on the annotation process are given in the following subsection.
5. The gastroenterologist completed part of the associated metadata:
    a. Number of polyps of interest found during the procedure;
    b. Current polyp ID;
    c. Polyp size, in millimetres;
    d. Paris classification [26,27];
    e. NICE classification [29];
    f. Preliminary diagnosis.
6. The pathologist completed part of the associated metadata:
    a. Final diagnosis;

b. Histological stratification.

7. Clinical metadata were exported into a csv file.

The protocol and all experiments comply with current Spanish and European Union legal regulations. The Basque Biobank was the source of samples and data. Each patient signed a specific document that was approved by the Ethics Committee of the Basque Country (CEIm-E) with identification code: PI+CES-BIOEF 2017-03.

2.1.2. Annotation Protocol

In order to obtain the annotated dataset, a systematic procedure was established (Figure 1):

1. Video clips were processed to extract the individual frames. Uninformative frames were discarded from this process and they includeD (Figure S1):

   a. Frames outside the patient;
   b. Blurry frames;
   c. Frames with high presence of bubbles;
   d. Frames with high presence of stool;
   e. Transition frames between WL and NBI.

2. All frames were analysed and categorized into showing a polyp or not, as well as identifying the type of light source (WL or NBI).
3. One out of 25 polyp frames (i.e., one frame per second) was selected to be manually annotated.
4. A researcher prepared three equally distributed sets of images to be processed using GTCreator [21]. Each set was manually annotated by one independent expert gastroenterologist with more than 15,000–25,000 (Á.J.C.; F.P.) and 500 (B.G.) colonoscopies. Furthermore, a void mask was also generated to indicate the valid endoscopic area of the image.
5. Segmented frames were collected and revised by a researcher to check completeness of the dataset prior to its use.
6. Manually segmented masks were automatically corrected with the void mask to adjust segmentations to the endoscopic image area.

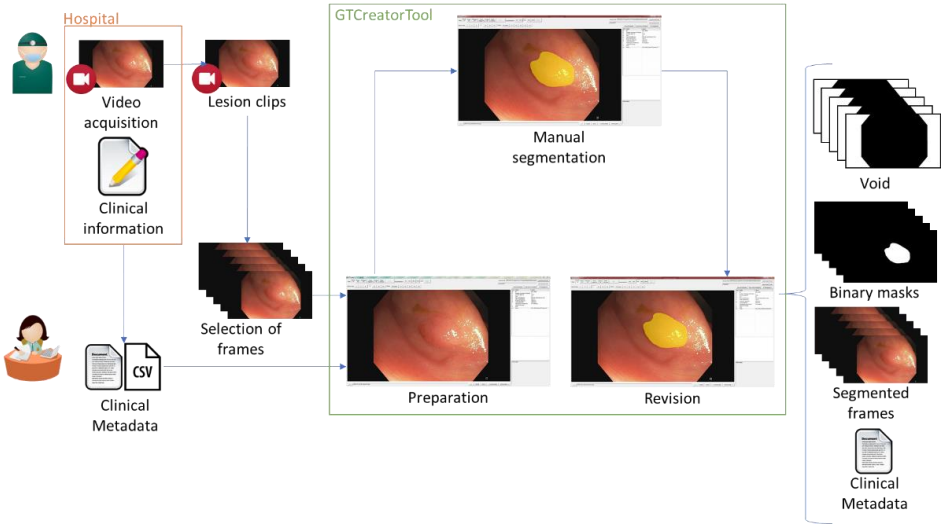


Figure 1. Annotation protocol for the PICCOLO dataset generation.

2.1.3. PICCOLO Dataset Details

Lesions were recorded between October 2017 and December 2019 at Hospital Universitario Basurto (Bilbao, Spain) using Olympus endoscopes (CF-H190L and-CF-HQ190L). Videos were recorded with an AVerMedia video capture and a hard drive storage. In all, the PICCOLO dataset included 76 lesions from 40 patients. In total, 62 out of these 76 lesions included white light (WL) and narrow band imaging (NBI) frames, while the remaining 14 lesions are recorded only using WL. Original videos were of length equal to 70.05 ± 59.28 s (range: 6–345 s), which corresponds to 1965.49 ± 1677.57 frames per clip (range: 187–10,364 frames). In all, more than 145,000 frames were revised. Out of the 80,847 frames showing a polyp, 3433 frames were selected for manual segmentation: 2131 WL images and 1302 NBI images (Table 1).

Table 1. Lesions and frames in the PICCOLO dataset according to clinical metadata.

Category	Items	# Lesions	# WL Frames	# NBI Frames
Paris Classification	Protruded lesions: 0-Ip	10	191	193
	Protruded lesions: 0-Ips	7	193	107
	Protruded lesions: 0-Is	16	387	253
	Flat elevated lesions: 0-IIa	29	928	436
	Flat elevated lesions:0-IIa/c	4	83	91
	Flat lesions: 0-IIb	2	85	21
	N/A	8	264	201
NICE classification	Type 1	17	404	284
	Type 2	50	1454	782
	Type 3	8	264	201
	N/A	1	9	35
Diagnosis	Adenocarcinoma	8	264	201
	Adenoma	50	1454	782
	Hyperplasia	17	404	284
	N/A	1	9	35
Histological stratification	High grade dysplasia	12	409	168
	Hyperplasia	13	312	225
	Invasive adenocarcinoma	8	264	201
	Low grade dysplasia	1	13	16
	No dysplasia	41	1124	657
	N/A	1	9	35

Image resolution are either 854 × 480 or 1920 × 1080, depending on the video they are acquired from. For each frame, three images are provided (Figure 2):

- Frame itself: png files showing the WL or NBI image.
- Mask: Binary mask indicating the area corresponding to the lesion.
- Void: Binary mask indicating the black area of the image.

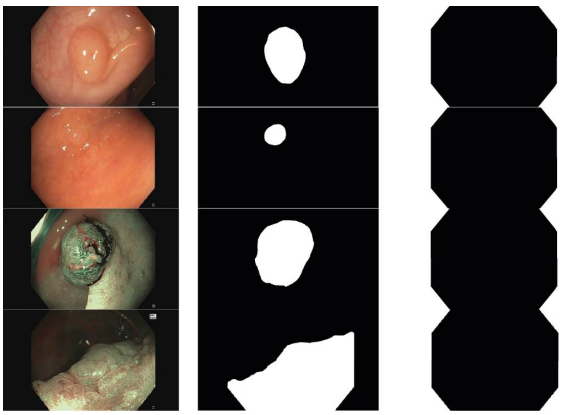


Figure 2. Examples of images and masks from the PICCOLO dataset. First two rows correspond to

white light images, while last two rows are NBI images. First column corresponds to the polyp frame, second column corresponds to the binary mask for the polyp area, and third column corresponds to the binary mask for the void area.

In the binary masks, pixels labelled with 1 correspond to the class (lesion or void area), and 0 otherwise (Figure 2).

Acquired images have been assigned to the train, validation or test set. In order to assure patient independence between sets, lesions originated from the same video are assigned to the same set. In all, there are 2203 (64.17%), 897 (26.13%) and 333 (9.70%) images in the train, validation and test set, respectively. Their clinical information is provided in Table 2. Furthermore, Table S1 in Supplementary Material provides all clinical information and assigned set for each of the 76 lesions.

Table 2. Frames in each of the sets according to clinical metadata.

Category	Items	Train Set	Validation Set	Test Set
Image type	WL	1382	558	192
	NBI	821	340	141
Paris Classification	Protruded lesions: 0-Ip	274	81	29
	Protruded lesions: 0-Ips	245	41	14
	Protruded lesions: 0-Is	433	176	31
	Flat elevated lesions: 0-IIa	1052	263	49
	Flat elevated lesions:0-IIa/c	27	122	25
	Flat lesions: 0-IIb	-	48	58
	N/A	172	166	127
NICE classification	Type 1	435	139	114
	Type 2	1552	592	92
	Type 3	172	166	127
	N/A	44	-	-
Diagnosis	Adenocarcinoma	172	166	127
	Adenoma	1552	592	92
	Hyperplasia	435	139	114
	N/A	44	-	-
Histological stratification	High grade dysplasia	360	217	-
	Hyperplasia	342	139	56
	Invasive adenocarcinoma	172	166	127
	Low grade dysplasia	-	-	29
	No dysplasia	1285	375	121
	N/A	44	-	-

The PICCOLO dataset is publicly available at the website of the Basque Biobank (https://www.biobancovasco.org/en/Sample-and-data-catalog/Databases/PD178-PICCOLO-EN.html), although a dedicated form to request downloading must be completed (the use of the dataset is restricted to research and educational purposes and use for commercial purposes is forbidden without prior written permission).

2.2. Public Datasets

In order to establish the utility of our dataset and range the learning capabilities of models trained over the present dataset, the other two publicly available datasets have been used in this work:

1. CVC-EndoSceneStill [23]. It contains 912 WL images which are manually segmented. A distribution into training, validation and test sets is provided by the owners. Each set contains 547, 183 and 182 images, respectively. This dataset is available at http://www.cvc.uab.es/CVC-Colon/index.php/databases/cvc-endoscenestill/.
2. Kvasir-SEG [24]. It contains 1000 WL images which are manually segmented. Distribution into training, validation and test sets has been randomly done, so each set includes 600, 200 and 200 images, respectively. This dataset is available at https://datasets.simula.no/kvasir-seg/.

These datasets do not provide clinical information of the images they include. In both cases, for each image, a binary mask indicating the polyp area is provided. While CVC-EndoSceneStill also provides void area binary masks, Kvasir-SEG does not. The polyp binary mask is used at training and testing, but the void binary mask is used to report metrics with respect to the valid endoscopic image, so void binary masks have been manually created for the images included in the test set of Kvasir-SEG.

2.3. Architectures and Training Process

In this study, we consider four models (Figure 3) which are obtained by combining a backbone (VGG-16 [30] or Densenet121 [31]) and a encoder–decoder architecture (U-Net [32] or LinkNet [33]), replicating the architectures and training parameters defined by Sánchez-Peralta et al. [34]. For implementation, segmentations models [35], Keras [36] and Tensorflow [37] have been used. Each model has been independently trained with the train and validation sets of each of the three datasets.

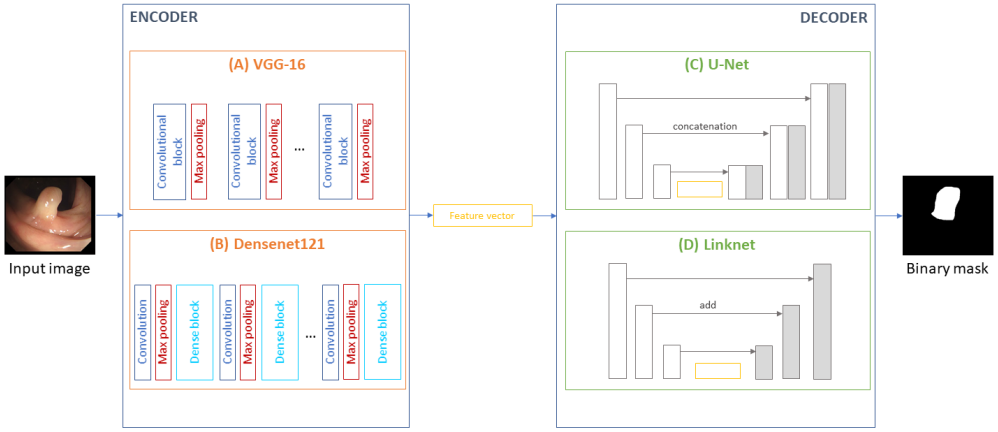


Figure 3. Models considered in this work are obtained by combining one backbone for the encoder (**A**,**B**) with one encoder-decoder architecture (**C**,**D**). Image reproduced from [34].

2.4. Reporting

In order to properly compare results, they are always reported over the test set of each dataset, thus over 182, 200 and 333 images for CVC-EndoSceneStill, Kvasir-SEG and PICCOLO, respectively. Furthermore, results are also reported over the whole test images (i.e., 715 images).

In order to characterize the images in each dataset, the following objective measures are calculated on each of the test sets:

1. Percentage of the image that corresponds to void area;
2. Percentage relative to the valid area that is polyp;
3. Mean value of the brightness channel in HSV [38], in the range [0, 1];
4. Histogram flatness measure [39], in the range [0, 1];
5. Histogram spread [39], in the range [0, 1].

For each test image, seven metrics are calculated: accuracy, precision, recall, specificity, F2-score, Jaccard index and Dice index, all of them based on the elements of the confusion matrix [34]. Only the valid area of the image (indicated in the void binary mask) is considered.

Appl. Sci. **2020**, *10*, 8501

3. Results and Discussion

3.1. Characterization of the Datasets

Table 3 shows the objective measures to characterize the datasets. Largest polyps are found in PICCOLO, but also with greater variability of sizes. Brightest and highest contrast images can be found in Kvasir-SEG, while CVC-EndoSceneStill and PICCOLO show similar appearance in these aspects.

Table 3. Characterization of the datasets. Results are reported as mean ± standard deviation. Minimum and maximum values are indicated between brackets. The void area refers to the black area in the images (Figure 2), while the remaining area is considered as valid area.

	CVC-EndoSceneStill	Kvasir-SEG	PICCOLO
Void area (%)	23.73 ± 5.57 (27.83 − 14.62)	15.23 ± 4.82 (28.44 − 6.16)	34.14 ± 0.33 (34.33 − 33.20)
Polyp area relative to the valid area (%)	12.50 ± 11.49 (66.15 − 0.75)	17.36 ± 15.65 (83.66 − 0.61)	20.45 ± 20.68 (86.11 − 0.00)
Mean value of brightness channel in HSV	0.560 ± 0.006 (1.000 − 0.000)	0.622 ± 0.003 (1.000 − 0.000)	0.532 ± 0.078 (1.000 − 0.000)
Histogram flatness measure	0.858 ± 0.121 (0.959 − 0.000)	0.419 ± 0.443 (0.962 − 0.000)	0.855 ± 0.215 (0.970 − 0.000)
Histogram spread	0.252 ± 0.088 (0.520 − 0.076)	0.218 ± 0.070 (0.432 − 0.075)	0.214 ± 0.086 (0.452 − 0.066)

It is also important to remark that the PICCOLO dataset includes WL and NBI images, which results in lower values in the brightness channel at the same time that the contrast is increased (higher values for the histogram flatness measure and lower values for histogram spread).

3.2. Comparison of Models Performance

Results of the Jaccard index for the different experiments are shown in Table 4. Tables with all metrics are provided in the Supplementary Material (Tables S2–S5). When reporting over the test set in CVC-EndoSceneStill, the best results are distributed between models trained with that same dataset and models trained with Kvasir-SEG. On the other hand, results in Kvasir-SEG and PICCOLO test sets are, in all cases but one, obtained when models are trained with the same dataset. Lastly, if all test sets are gathered and analysed together as whole, the best performance is obtained with models trained using the PICCOLO dataset. Models trained with Kvasir-SEG show better performance than those trained with CVC-EndoSceneStill. Therefore, it can be observed that the PICCOLO dataset allows for training more generalized models, that perform more uniformly along all datasets, rather than obtaining the best results for its own test set.

Table 4. Jaccard index for each experiment. Best value per train/validation set and test set is indicated in bold.

Dataset for Training/Validation Set	Network	Dataset for Test Set			
		CVC-EndoSceneStill	Kvasir-SEG	PICCOLO	All
CVC-EndoSceneStill	U-Net + VGG16	56.80 ± 36.09	46.29 ± 29.36	32.82 ± 34.81	42.69 ± 35.13
	U-Net + Densenet121	67.31 ± 34.47	56.12 ± 34.29	38.76 ± 39.05	50.88 ± 38.51
	LinkNet + VGG16	62.77 ± 33.81	51.13 ± 27.67	30.55 ± 33.57	44.51 ± 34.86
	LinkNet + Densenet121	**72.16 ± 30.93**	**56.69 ± 33.68**	**39.52 ± 37.90**	**52.63 ± 37.53**
Kvasir-SEG	U-Net + VGG16	32.44 ± 38.79	64.23 ± 29.78	28.51 ± 35.12	39.50 ± 37.98
	U-Net + Densenet121	**71.82 ± 29.87**	74.13 ± 23.40	44.78 ± 38.73	**59.87 ± 35.72**
	LinkNet + VGG16	58.62 ± 36.18	72.53 ± 23.92	40.43 ± 36.41	54.04 ± 35.99
	LinkNet + Densenet121	69.10 ± 32.53	**74.52 ± 22.81**	**44.92 ± 37.37**	59.35 ± 35.33
PICCOLO	U-Net + VGG16	47.76 ± 36.46	52.64 ± 30.41	58.74 ± 36.06	54.24 ± 34.93
	U-Net + Densenet121	62.41 ± 34.78	**65.33 ± 30.66**	**64.01 ± 36.23**	**63.97 ± 34.35**
	LinkNet + VGG16	54.58 ± 35.63	58.99 ± 30.40	54.46 ± 38.88	55.76 ± 35.87
	LinkNet + Densenet121	**64.18 ± 33.04**	59.61 ± 33.80	60.14 ± 38.31	61.02 ± 35.79

If comparison is done over the different test sets, the lowest results are obtained in the PICCOLO dataset, which might be because it contains a wider range of different polyps which makes it more difficult to learn with any of the used datasets. This result might be due to the presence of bias within each dataset, which would lead to the recommendation of multi-centre or cross-dataset evaluation for DL methods in order to properly understand the performance of such models in a real-world settings [40].

Similar to the results found in previous works [34], it can be concluded that in general terms, Densenet121 works better than VGG-16 as backbone and the LinkNet encoder–decoder architecture obtains better results than U-Net.

Sets independence is assured in our dataset, as all lesions obtained from the same patient are assigned to the same dataset. This is also considered when determining train, validation and test sets in CVC-EndoSceneStill [23]. On the other hand, Kvasir-SEG does not indicate the origin of each image nor provide a predefined division into sets, it is not possible to determine if the random allocation of images might have violated this independency. Should this happen, then the higher values for models trained and reported on Kvasir-SEG might be explained.

Regarding the clinical characteristics of the lesions in the test sets, it is not possible to establish a comparison as CVC-EndoSceneStill and Kvasir-SEG datasets do not provide clinical information.

Lastly, the inclusion of WL together with NBI images in the PICCOLO dataset might increase independence against changes in colour when training the models, showing better generalization when results are reported over the test sets of the other two datasets. In this regard, metrics have been calculated considering only the WL frames of the test set in the PICCOLO dataset for a fairer comparison for the other datasets. Results in the PICCOLO test sets remains similar regardless of the type of images considered (Table S4). When the three datasets are analysed together, including only the WL frames from PICCOLO (Table S5), the performances of Kvasir-SEG and PICCOLO are more similar.

3.3. Influence of the Polyp Morphology in the Results

Figure 4 shows the Jaccard index for the four analysed models, when polyps in the test set of the PICCOLO dataset are classified based on their morphology accordingly to the Paris classification. It can be clearly seen that flat polyps (0-IIb) obtain the lowest results in all models, and that pedunculated and sessile polyps (0-Ip, 0-Ips and 0-Is) outperform the flat ones (0-IIa, 0-IIa/c and 0-IIb). This result is well aligned with the clinical findings shown by clinical works [8,9]. Furthermore, it is important to remark that the best results for flat polyps are obtained with models trained with PICCOLO dataset in three out of the four considered models. Regretfully, this analysis cannot be done with models trained using CVC-EndoSceneStill and Kvasir-SEG, as polyp morphology is not available for those datasets. Similarly, Lee et al. [41] employed a proprietary dataset to train the polyp detection method based on YOLOv2, and they also found that DL methods show lower sensitivity for flat polyps.

3.4. Current Limitations and Future Work

This work has also some limitations that should be acknowledged. In the first place, flat polyps, specially 0-IIb, according to the Paris classification, are still underrepresented in the dataset, in common with other datasets. This type is less frequently identified, therefore a longer, multicentre acquisition campaign would be necessary to increase their presence in the public datasets. Furthermore, these polyps are also the most difficult to be detected due to their subtle appearance, thus the Computer Assisted Diagnosis (CAD) systems would be more helpful to assist their detection [42]. We consider that increasing the number of images showing these challenging polyps would increase the clinical utility of the CAD systems, although assistance for the rest of types would also remain useful for novice endoscopists with lower Adenoma Detection Rate (ADR). Secondly, images in the dataset are mainly "polyp centred". This means that the polyp is clearly visible in the image. This type of image is highly convenient for polyp segmentation as well as classification. Nevertheless, we consider that the inclusion of images where the lumen at or near the screen centre, leaving the polyps peripheral, a typical

setup during the clinical exploration, would be of help for detection systems. Besides, the availability of long video sequences, even if not all frames are manually segmented, would also benefit the development of such systems. Lastly, manual segmentation is prone to inter-observer variability in medical images [43,44], therefore uncertainty should be considered and analysed, both at the dataset creation stage, but it would also be interesting to take it into account the analysis of segmentation results [45].

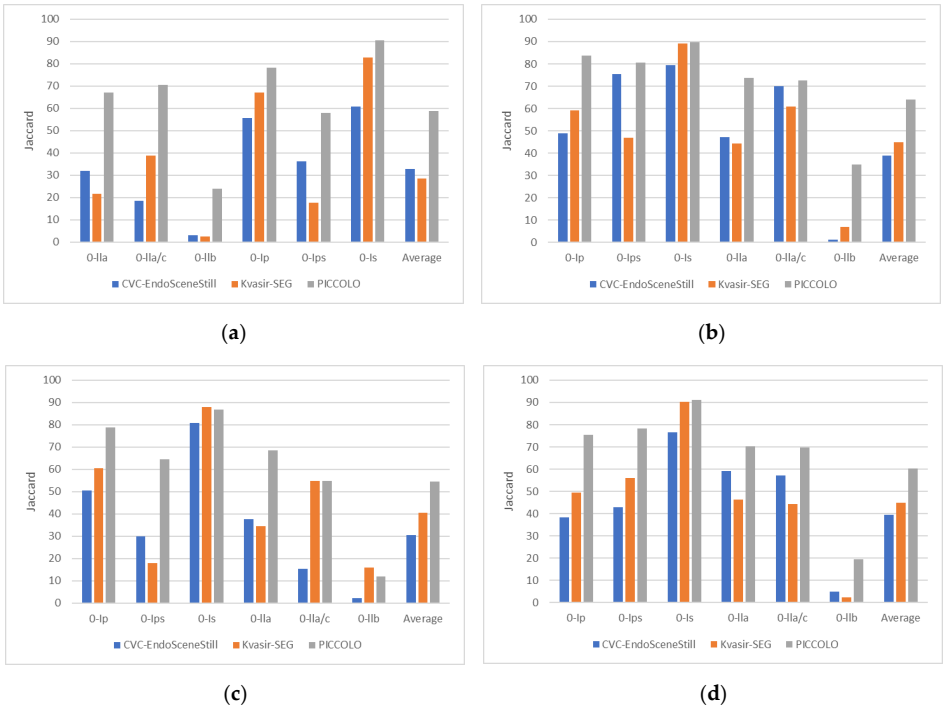


Figure 4. Jaccard index reported on the test set of the PICCOLO dataset for the four models. Each series represents the model trained with the corresponding dataset (**a**) U-Net + VGG16; (**b**) U-Net + Densenet121; (**c**) LinkNet + VGG16; (**d**) LinkNet + Densenet121.

As future work, we consider the expansion of the current dataset to include some of the previously mentioned images and sequences. A joint global initiative to gather all efforts would result in a larger and more diverse dataset of polyps, increasing the possibilities of the research community.

Lastly, it is worth mentioning that, after developing any deep learning method, it is essential to carry out prospective studies and randomized trials to compare their performance with experts clinicians, which is so far not well proven [46], with the final aim of increasing the adenoma detection rate, as indicator of CRC detection. In this regard, a survey is currently undergoing to compare the identification of polyps done by expert gastroenterologists and residents with a deep learning model trained with the PICCOLO dataset.

4. Conclusions

This work presents a new dataset for polyp detection, localisation and segmentation. It provides 3433 polyp images together with a manually annotated binary mask of the polyp area. It also provides a set of clinical metadata for each of the lesions included. Besides, we have compared four different models trained with our PICCOLO dataset and two other publicly available datasets (CVC-EndoSceneStill and Kvasir-SEG). Our results show that the PICCOLO dataset is suitable for training deep learning models for polyp segmentation, which results in better generalization capabilities

as well as better results for flat polyps. We also provide clinical metadata which, as far as the authors know, are not available in other publicly available datasets, and which might eventually increase the potential use of this dataset for polyp classification purposes.

Supplementary Materials: The following are available online at http://www.mdpi.com/2076-3417/10/23/8501/s1, Table S1: Clinical metadata and distribution of lesions into training, validation and test sets. Table S2: Metrics obtained in the CVC-EndoSceneStill test set for the different models and datasets used during training. Table S3: Metrics obtained in the Kvasir-SEG test set for the different models and datasets used during training. Table S4: Metrics obtained in the PICCOLO test set for the different models and datasets used during training. Table S5: Metrics obtained for all test sets as a whole for the different models and datasets used during training. Figure S1: Uninformative frames.

Author Contributions: Conceptualization, L.F.S.-P., J.B.P. and A.P.; methodology, L.F.S.-P., J.B.P., A.P., N.A., R.B. and C.L.S.; software, L.F.S.-P., and A.P.; validation, L.F.S.-P., J.B.P. and A.P.; formal analysis, L.F.S.-P.; investigation, L.F.S.-P., Á.J.C., F.P., N.A., R.B. and B.G.; resources, J.B.P., A.P., R.B., C.L.S., and F.M.S.-M.; data curation, L.F.S.-P., Á.J.C., F.P., N.A. and B.G.; writing—original draft preparation, L.F.S.-P. and J.B.P.; writing—review and editing, L.F.S.-P., J.B.P., A.P., Á.J.C., F.P., N.A., R.B., B.G., C.L.S., and F.M.S.-M.; visualization, L.F.S.-P.; supervision, J.B.P., A.P. and F.M.S.-M.; project administration, J.B.P., A.P. and C.L.S.; funding acquisition, J.B.P., A.P., R.B., C.L.S. and F.M.S.-M. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partially supported by PICCOLO project. This project has received funding from the European Union's Horizon2020 research and innovation programme under grant agreement No 732111. The sole responsibility of this publication lies with the author. The European Union is not responsible for any use that may be made of the information contained therein. Furthermore, this publication has also been partially supported by GR18199 from Consejería de Economía, Ciencia y Agenda Digital of Junta de Extremadura (co-funded by European Regional Development Fund–ERDF. "A way to make Europe"/ "Investing in your future". This work has been performed by the ICTS "NANBIOSIS" at the Jesús Usón Minimally Invasive Surgery Centre.

Acknowledgments: Authors want to thank for their invaluable contribution to the data acquisition process to the pathologists from Hospital Universitario Basurto: Nagore Arbide, Jaques Velasco and Maria Carmen Etxezárraga; Ainara Egia and technical staff from Basque Biobank, Basque Foundation for Health Innovation and Research-BIOEF; as well as Virginia Cabezón and Agueda Azpeitia.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. International Agency for Research on Cancer. *Colorectal Cancer Factsheet*; International Agency for Research on Cancer: Lyon, France, 2018.
2. World Health Organization. *World Cancer Report 2014*; Stewart, B.W., Wild, C.P., Eds.; International Agency for Research on Cancer: Lyon, France, 2014; ISBN 978-92-832-0429-9.
3. Siegel, R.L.; Miller, K.D.; Jemal, A. Cancer statistics, 2015. *CA Cancer J. Clin.* **2015**, *65*, 29. [CrossRef] [PubMed]
4. Siegel, R.L.; Miller, K.D.; Jemal, A. Cancer statistics, 2020. *CA Cancer J. Clin.* **2020**, *70*, 7–30. [CrossRef] [PubMed]
5. Wiegering, A.; Ackermann, S.; Riegel, J.; Dietz, U.A.; Götze, O.; Germer, C.T.; Klein, I. Improved survival of patients with colon cancer detected by screening colonoscopy. *Int. J. Colorectal Dis.* **2016**, *31*, 1039–1045. [CrossRef] [PubMed]
6. Kaminski, M.F.; Thomas-Gibson, S.; Bugajski, M.; Bretthauer, M.; Rees, C.J.; Dekker, E.; Hoff, G.; Jover, R.; Suchanek, S.; Ferlitsch, M.; et al. Performance measures for lower gastrointestinal endoscopy: A European Society of Gastrointestinal Endoscopy (ESGE) quality improvement initiative. *United Eur. Gastroenterol. J.* **2017**, *5*, 309–334. [CrossRef]
7. Lund, M.; Trads, M.; Njor, S.H.; Erichsen, R.; Andersen, B. Quality indicators for screening colonoscopy and colonoscopist performance and the subsequent risk of interval colorectal cancer: A systematic review. *JBI Database Syst. Rev. Implement. Reports* **2019**. [CrossRef]
8. Kim, N.H.; Jung, Y.S.; Jeong, W.S.; Yang, H.-J.; Park, S.-K.; Choi, K.; Park, D. Il Miss rate of colorectal neoplastic polyps and risk factors for missed polyps in consecutive colonoscopies. *Intest. Res.* **2017**, *15*, 411. [CrossRef]
9. Klare, P.; Sander, C.; Prinzen, M.; Haller, B.; Nowack, S.; Abdelhafez, M.; Poszler, A.; Brown, H.; Wilhelm, D.; Schmid, R.M.; et al. Automated polyp detection in the colorectum: A prospective study (with videos). *Gastrointest. Endosc.* **2019**, *89*, 576–582. [CrossRef]

10. Ishaq, S.; Siau, K.; Harrison, E.; Tontini, G.E.; Hoffman, A.; Gross, S.; Kiesslich, R.; Neumann, H. Technological advances for improving adenoma detection rates: The changing face of colonoscopy. *Dig. Liver Dis.* **2017**, *49*, 721–727. [CrossRef]
11. Byrne, M.F.; Shahidi, N.; Rex, D.K. Will Computer-Aided Detection and Diagnosis Revolutionize Colonoscopy? *Gastroenterology* **2017**, *153*, 1460–1464. [CrossRef]
12. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef]
13. Kim, J.; Hong, J.; Park, H. Prospects of deep learning for medical imaging. *Precis. Futur. Med.* **2018**, *2*, 37–52. [CrossRef]
14. Shen, D.; Wu, G.; Suk, H.-I. Deep Learning in Medical Image Analysis. *Annu. Rev. Biomed. Eng.* **2017**, *19*, 221–248. [CrossRef] [PubMed]
15. Chan, H.-P.; Samala, R.K.; Hadjiiski, L.M.; Zhou, C. Deep Learning in Medical Image Analysis. In *Deep Learning in Medical Image Analysis*; Lee, G., Fujita, H., Eds.; Springer: Berlin/Heidelberg, Germany, 2020.
16. Aziz, M.; Fatima, R.; Dong, C.; Lee-Smith, W.; Nawras, A. The impact of deep convolutional neural network-based artificial intelligence on colonoscopy outcomes: A systematic review with meta-analysis. *J. Gastroenterol. Hepatol.* **2020**, 1–8. [CrossRef]
17. Sánchez-Peralta, L.F.; Bote-Curiel, L.; Picón, A.; Sánchez-Margallo, F.M.; Pagador, J.B. Deep learning to find colorectal polyps in colonoscopy: A systematic literature review. *Artif. Intell. Med.* **2020**, *108*. [CrossRef] [PubMed]
18. Nogueira-Rodríguez, A.; Domínguez-Carbajales, R.; López-Fernández, H.; Iglesias, Á.; Cubiella, J.; Fdez-Riverola, F.; Reboiro-Jato, M.; Glez-Peña, D. Deep Neural Networks approaches for detecting and classifying colorectal polyps. *Neurocomputing* **2020**. [CrossRef]
19. Min, J.K.; Kwak, M.S.; Cha, J.M. Overview of deep learning in gastrointestinal endoscopy. *Gut Liver* **2019**, *13*, 388–393. [CrossRef]
20. Borgli, H.; Thambawita, V.; Smedsrud, P.; Hicks, S.; Jha, D.; Eskeland, S.; Randel, K.R.; Pogorelov, K.; Lux, M.; Dang-Nguyen, D.-T.; et al. HyperKvasir: A Comprehensive Multi-Class Image and Video Dataset for Gastrointestinal Endoscopy. *Sci. Data* **2020**, *7*. [CrossRef]
21. Bernal, J.; Histace, A.; Masana, M.; Angermann, Q.; Sánchez-Montes, C.; de Miguel, C.R.; Hammami, M.; García-Rodríguez, A.; Córdova, H.; Romain, O. GTCreator: A flexible annotation tool for image-based datasets. *Int. J. Comput. Assist. Radiol. Surg.* **2019**, *14*, 191–201. [CrossRef]
22. Tajbakhsh, N.; Jeyaseelan, L.; Li, Q.; Chiang, J.N.; Wu, Z.; Ding, X. Embracing Imperfect Datasets: A Review of Deep Learning Solutions for Medical Image Segmentation. *Med. Image Anal.* **2020**, *63*, 101693. [CrossRef]
23. Vázquez, D.; Bernal, J.; Sánchez, F.J.; Fernández-Esparrach, G.; López, A.M.; Romero, A.; Drozdzal, M.; Courville, A. A Benchmark for Endoluminal Scene Segmentation of Colonoscopy Images. *J. Healthc. Eng.* **2017**. [CrossRef]
24. Jha, D.; Smedsrud, P.H.; Riegler, M.A.; Halvorsen, P.; de Lange, T.; Johansen, D.; Johansen, H.D. Kvasir-SEG: A Segmented Polyp Dataset. In *Proceedings of the International Conference on Multimedia Modeling (MMM)*; Springer: Cham, Switzerland, 2020.
25. Angermann, Q.; Bernal, J.; Sánchez-Montes, C.; Hammami, M.; Fernández-Esparrach, G.; Dray, X.; Romain, O.; Sánchez, F.J.; Histace, A. Towards Real-Time Polyp Detection in Colonoscopy Videos: Adapting Still Frame-Based Methodologies for Video Sequences Analysis. In *Computer Assisted and Robotic Endoscopy and Clinical Image-Based Procedures*; Springer: Cham, Switzerland, 2017; pp. 29–41. ISBN 978-3-319-67542-8.
26. Participants in the Paris Workshop The Paris endoscopic classification of superficial neoplastic lesions: Esophagus, stomach, and colon. *Gastrointest. Endosc.* **2003**, *58*, S3–S43. [CrossRef]
27. Endoscopic Classification Review Group Update on the Paris Classification of Superficial Neoplastic Lesions in the Digestive Tract. *Endoscopy* **2005**, *37*, 570–578. [CrossRef] [PubMed]
28. Mesejo, P.; Pizarro, D.; Abergel, A.; Rouquette, O.; Beorchia, S.; Poincloux, L.; Bartoli, A. Computer-Aided Classification of Gastrointestinal Lesions in Regular Colonoscopy. *IEEE Trans. Med. Imaging* **2016**, *35*, 2051–2063. [CrossRef] [PubMed]
29. Hattori, S.; Iwatate, M.; Sano, W.; Hasuike, N.; Kosaka, H.; Ikumoto, T.; Kotaka, M.; Ichiyanagi, A.; Ebisutani, C.; Hisano, Y.; et al. Narrow-band imaging observation of colorectal lesions using NICE classification to avoid discarding significant lesions. *World J. Gastrointest. Endosc.* **2014**, *6*, 600. [CrossRef]

30. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
31. Huang, G.; Liu, Z.; van der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. *arXiv* **2018**. [CrossRef]
32. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015. Lecture Notes in Computer Science*; Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F., Eds.; Springer: Berlin/Heidelberg, Germany, 2015; Volume 9351, pp. 234–241. ISBN 9783319245737.
33. Chaurasia, A.; Culurciello, E. LinkNet: Exploiting encoder representations for efficient semantic segmentation. *2017 IEEE Vis. Commun. Image Process. VCIP* **2018**, 1–4. [CrossRef]
34. Sánchez-Peralta, L.F.; Picón, A.; Antequera-Barroso, J.A.; Ortega-Morán, J.F.; Sánchez-Margallo, F.M.; Pagador, J.B. Eigenloss: Combined PCA-Based Loss Function for Polyp Segmentation. *Mathematics* **2020**, *8*, 1316. [CrossRef]
35. Yakubovskiy, P. Segmentation Models. Available online: https://github.com/qubvel/segmentation_models (accessed on 24 November 2020).
36. Chollet, F. Keras. Available online: https://github.com/keras-team/keras (accessed on 24 November 2020).
37. Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G.S.; Davis, A.; Dean, J.; Devin, M.; et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *CoRR* **2015**. [CrossRef]
38. Abaza, A.; Harrison, M.A.; Bourlai, T. Quality metrics for practical face recognition. In Proceedings of the Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012), Tsukuba, Japan, 11–15 November 2012; pp. 3103–3107.
39. Tripathi, A.K.; Mukhopadhyay, S.; Dhara, A.K. Performance metrics for image contrast. In Proceedings of the 2011 International Conference on Image Information Processing, Shimla, India, 3–5 November 2011; pp. 1–4.
40. Thambawita, V.; Jha, D.; Hammer, H.L.; Johansen, H.D.; Johansen, D.; Halvorsen, P.; Riegler, M.A. An Extensive Study on Cross-Dataset Bias and Evaluation Metrics Interpretation for Machine Learning Applied to Gastrointestinal Tract Abnormality Classification. *ACM Trans. Comput. Healthc.* **2020**, *1*, 1–29. [CrossRef]
41. Lee, J.Y.; Jeong, J.; Song, E.M.; Ha, C.; Lee, H.J.; Koo, J.E.; Yang, D.H.; Kim, N.; Byeon, J.S. Real-time detection of colon polyps during colonoscopy using deep learning: Systematic validation with four independent datasets. *Sci. Rep.* **2020**, *10*, 1–9. [CrossRef]
42. Sánchez-Peralta, L.F.; Sánchez-Margallo, F.M.; Bote Chacón, J.; Soria Gálvez, F.; López-Saratxaga, C.; Picón Ruiz, A.; Pagador, J.B. Is it necessary to improve the colorectal polyps databases for detection CAD systems based on deep learning? *Br. J. Surg.* **2018**, *105*, 5–14.
43. Joskowicz, L.; Cohen, D.; Caplan, N.; Sosna, J. Inter-observer variability of manual contour delineation of structures in CT. *Eur. Radiol.* **2019**, *29*, 1391–1399. [CrossRef] [PubMed]
44. Chlebus, G.; Meine, H.; Thoduka, S.; Abolmaali, N.; Van Ginneken, B.; Hahn, H.K.; Schenk, A. Reducing inter-observer variability and interaction time of MR liver volumetry by combining automatic CNN-based liver segmentation and manual corrections. *PLoS ONE* **2019**, *14*, e0217228. [CrossRef] [PubMed]
45. Wickstrøm, K.; Kampffmeyer, M.; Jenssen, R. Uncertainty and Interpretability in Convolutional Neural Networks for Semantic Segmentation of Colorectal Polyps. *Med. Image Anal.* **2020**, *60*. [CrossRef] [PubMed]
46. Nagendran, M.; Chen, Y.; Lovejoy, C.A.; Gordon, A.C.; Komorowski, M.; Harvey, H.; Topol, E.J.; Ioannidis, J.P.A.; Collins, G.S.; Maruthappu, M. Artificial intelligence versus clinicians: Systematic review of design, reporting standards, and claims of deep learning studies in medical imaging. *BMJ* **2020**, *368*, 1–12. [CrossRef] [PubMed]

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).

A Material suplementario

A.1 Medical needs related to the endoscopic technology and colonoscopy for colorectal cancer diagnosis

El material suplementario del artículo *Medical needs related to the endoscopic technology and colonoscopy for colorectal cancer diagnosis* está compuesto de los siguientes documentos:

A.1.1 - Guía de la entrevista
A.1.2 - Cuestionario final
A.1.3 - Citas literales de los participantes
A.1.4 - Detalles de los resultados de la asignación de Dirichlet latente

A.1.1 Guía de la entrevista

Additional File 1

- **Block 1. Presentation (1-2 minutes)**

1. Self-introduction of the interviewer

2. Presentation of the project

3. Objective of the study

4. Outline of the interview:

    - Block 2: demographic data

    - Block 3: current procedure and technologies for colonoscopy and colorectal cancer diagnosis

    - Block 4: challenges to be addressed to improve the current procedure and technologies

5. Acceptance to voluntarily participate in the interview

6. Permission for recording

- **Block 2. Demography (up to 5 minutes)**

1. Please, make a briefly introduce yourself: name, age, academic training, number of colonoscopies performed, where you are working or have worked, in what position, etc.

- **Block 3. Current procedure for colonoscopy and colorectal cancer diagnosis (10 minutes)**

1. What is the current screening procedure for the diagnosis of colorectal cancer?

2. But more specifically, and more related to your field of work, how is the current colonoscopy procedure?

3. What problems do you encounter while performing such procedure? And related to the technology used?

- **Block 4. Challenges (15 minutes)**

1. What improvements would you like to see in the colonoscopy procedure? And related to current technology and equipment (expected innovations)?

2. What is your opinion and personal experience of the use of advanced imaging techniques (including dye-spray, virtual chromoendoscopy, photonics techniques) for the diagnosis of polyps and colorectal cancer?

3. If you could have at your disposal an ideal device for supporting your decision-making in the assessment of polyps or colorectal cancer, what additional information would you like to be provided with? (Detection, classification, visual information, auditory information…)

A.1.2 Cuestionario final

Additional File 2

- **Block 1. Demographics**

1. In what country do you work?

2. Please indicate your area of practice.

- ☐ Trainee colonoscopist
- ☐ Qualified colonoscopist
- ☐ Qualified colonoscopist with a special interest in cancer screening
- ☐ Qualified colonoscopist with a special interest in therapeutic endoscopy

3. During your professional career, how many colonoscopies have you performed?

- ☐ Less than 50 procedures
- ☐ 50-250 procedures
- ☐ 250-1000 procedures
- ☐ More than 1000 procedures

- **Block 2. Detection of colorectal lesions**

4. If a software program were available to help you detecting polyps, for which type would you find it more useful? Please order from more to less usefulness according to Paris classification (1 - more usefulness; 5 - less usefulness)

- ☐ 0-Ip - Pedunculated
- ☐ 0-Is - Sessile
- ☐ 0-IIa - Flat elevated
- ☐ 0-IIb - Completely flat
- ☐ 0-IIc - Slightly depressed

5. If you had access to computer assisted diagnostic (CAD) programs, how would you prefer to be alerted to the detection of a polyp?

- ☐ Audible alarm
- ☐ Visual cue on screen (away from the endoscopic image)
- ☐ Highlighting the polyp on the endoscopic image (augmented display)
- ☐ Other (please specify)

6. If you have further comments or insights about a CAD system for polyp detection, please indicate them below.

- **Block 3. Characterisation of colorectal lesions**

7. According to your experience, what is more helpful to establish the malignancy of a lesion? Please order from more to less helpfulness.

- ☐ Paris classification
- ☐ Vascularity
- ☐ Lession surface (granularity / no granularity)

- ☐ Kudo's pit pattern
- ☐ Other

8. If you have chosen "other" in the previous question, please explain here in what you base the malignancy of the lesion.

9. Do you think the existing methods to establish the malignancy of a lesion are simple and reproducible enough?

- ☐ Yes
- ☐ No

10. Would you like to have an automated system that would give you a diagnosis, without having to make an interpretation, of the lesion you are facing (neoplastic, non-neoplastic, submucosal invasion)?

- ☐ Yes
- ☐ No

11. How would you like to receive the feedback on the diagnosis suggestion?

- ☐ Audible alarm
- ☐ Traffic light cues (Green – benign, Yellow – pre-malign, Red – malign)
- ☐ Text
- ☐ Other (please specify)

12. What information would you like to obtain?

- ☐ Diagnosis suggestions
- ☐ Grade
- ☐ Width
- ☐ Depth estimation
- ☐ Margin estimation
- ☐ Infiltration probability
- ☐ Other (please specify)

13. In the presence of large lesions that are not polypoid or with margins that are difficult to differentiate, could you completely remove the lesion with more confidence if you had a CAD system? (compared to what you are using now)

- ☐ Yes
- ☐ No

14. In the surveillance of the scar after having performed a piecemeal polypectomy, you look for residual neoplasia. Do you think that the CAD system could contribute to detect the residual lesion?

- ☐ Yes
- ☐ No

15. If you have further comments or insights about a CAD system for polyp diagnosis, please indicate them below.

A.1.3 Citas literales de los participantes

1 ***Additional File 3***

Need	**Most representative quotes from interviews**
Clinical needs	
To improve polyp detection, especially for flat polyps, to reduce the polyp missing rate	*"A large polyp, larger than 1 cm, is difficult for any endoscopist to miss. The polyps that are like a mushroom, you will be able to see them easily. The problem is with those that are flat, that is, it is a roughness in the area you are seeing. So that, if you do not see it clearly or there is not a system that allows you to locate it or be able to define it well, those are the ones that are going to give you more problems."*
	"There are different questions. The first one is to see them [all polyps] and not leave anything, and the other one is to take an attitude to a lesion that you already know."
To minimize the resection of hyperplastic polyps	*"Another limitation is having to remove all the polyps, being some polyps more dangerous and others that are not at all. This limitation forces us to remove the polyps with the risk for the patient and the expense of time and material that causes."*
In-situ classification of detected polyps to reduce costs and time	*"The shorter the waiting [for histology] is, subjectively, the better, you close an episode earlier, the patient is discharged."*
To reduce recurrence rate by assessing lesion margins and inspecting the remaining tissue	*"To be able to identify on a resection bed if any adenomatous residue remains. That is important."*
To assess the depth of submucosal infiltration for detected lesions	*"In the case of large polyps, these usually generate major doubts when you have to say whether there is invasion or not. And also, how deep the invasion reaches, because it is not the same reaching a part of the submucosa that can be completely removed with dissection or deeper. We are talking about microns."*
	"It is important to decide not touching this polyp because here there is a submucosal invasion, so the polypectomy to be done is going to take a risk and it will not cure the patient."
Computer Aided Diagnosis (CAD) system needs	
to support the polyp detection and diagnosis with additional information for a better diagnosis	*"I think it is also very difficult to interpret the images. We do not have a histological vision, like pathologists."*
	"The CAD system detects the polyp and tells you that it is a polyp and draws a square, and tells you that this polyp is adenomatous, or this polyp has an area that looks like a tumor, or this polyp is hyperplastic and you do not have to do anything with it because of the place where it is located. All those things, or that it locates it with more precision."
To provide information about physical characteristics of the polyps	*"The size [of polyps], because the measurements that are made now are estimates."*
	"A system that tells you there is a thing... it measures X, and this is hyperplastic. And then I can say in the report, at such distance there is a lesion of X millimeters, which is of hyperplastic

1

	characteristics."
To accurately visualize the polyp contour	*"I think highlighting the polyp from the rest of the mucosa would be important because sometimes you do not see very well where the edge of the polyp is. Then, if it could be delimited better by image, this would help at the time of doing the resection."*
To indicate polyp invasiveness or remaining adenomatous tissue left after resection	*"Of course, no manufacturer has a system that can indicate if what I see infiltrates submucosal and if you have left an adenomatous residue."*
To provide information in a visual manner	*"[That the system provides] visual [information] will always be better."*
Operational/physical needs	
Better lighting inside the colon	*"The colonoscopy itself already has its difficulties: visualizing between the folds, the angulations; there are essentially flat lesions or with morphology or appearance similar to the normal mucosa, which can often be missed, and which would be desirable to identify in greater numbers, more easily. On a day-to-day basis, the tests have limited time. Many times, you spend time in washing, in cleaning properly, and the time of exploration is reduced, and it would be a good thing [to minimize any delay due to this]."*
	"Another issue is the visualization of certain areas that are slightly dark. The colon has haustra, so you look straight ahead, but there are things that you eventually might not see. Those in front, you can see them, but those that are a little aside maybe you do not see them totally or with total clarity."
	"Both the image capture and the lighting inside the colon. In those two fields, I think that we can move forward."
Better visualization and exploration of the mucosa	*"If you can rotate the tip more in polyps that are more on the back side of the haustrum, you could work better. Now you do not have much rotation, the tip is not so flexible."*
Compromise between flexibility and stiffness of instruments	*"It is complicated because in addition the material that you have to pass through the working channel is semi-rigid. It has to have functionality from the outside. It is complicated. But that [being able to pass easily] would help."*
	"The problem is that if you have one [endoscope] with a lot of flexibility, you do not have enough rigidity to be able to transmit the movements to the tip in an area that you have focused. For example, when you access the colon with the gastro[scope] and you have to do many laps, eventually you do not have enough stiffness to be able to reach and transmit that force."
	"More flexible tips are being tried, variable stiffness. But I do not see much advantage to perform the colonoscopy better with variable stiffness. That is, they are devices with which you can reach your goal in more than a 95%, but depending on the patient's anatomy, it may take more or less effort".
Equipment with better image quality	*"We all have 4K televisions at home, and here we*

2

	are getting to high definition. So, I suppose it will be by miniaturization of the chips or whatever, but the imaging technology is much more advanced in household equipment, and here it has not arrived, I do not know why. I think that the more definition you have, the more you will get."
	"The problems of image quality, because we have mixed technology in terms of image qualities, we have high definition with no really high definition, and that is always a problem, which can sometimes be solved. You have technology, models of endoscopes with different levels of quality and image".

2 **Table 1. Verbatim quotes from participants in the interviews**

3

A.1.4 Detalles de los resultados de la asignación de Dirichlet latente

Additional File 4

Figure 1. Words of the Topic 1, which revolves around the *patient*

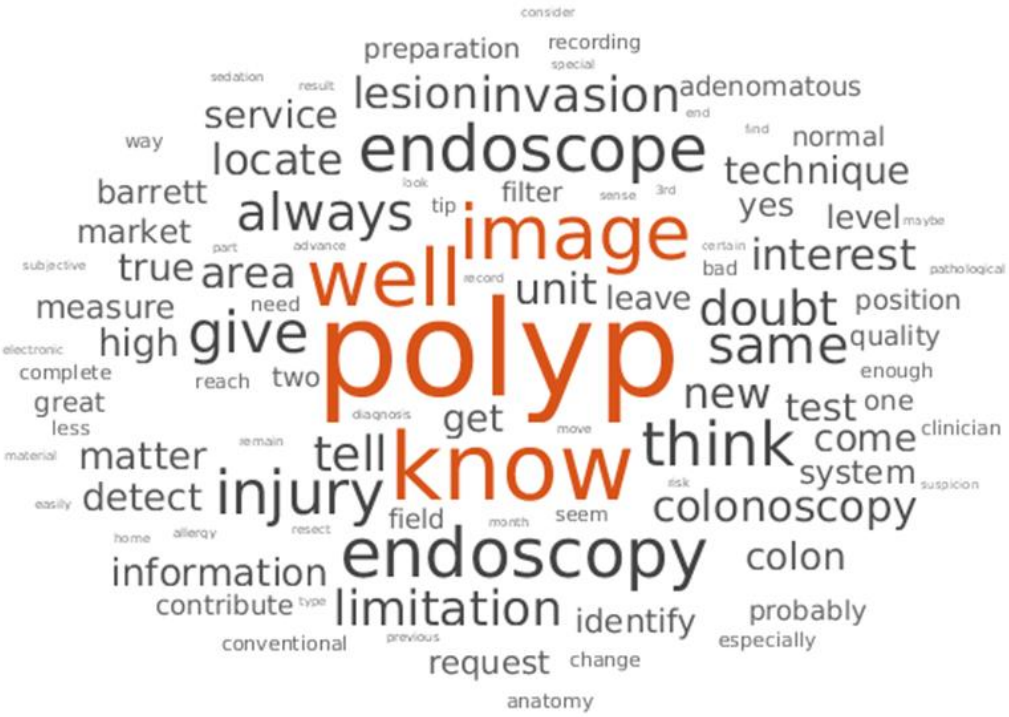


Figure 2. Words of the Topic 2, which revolves around the *polyp*

1

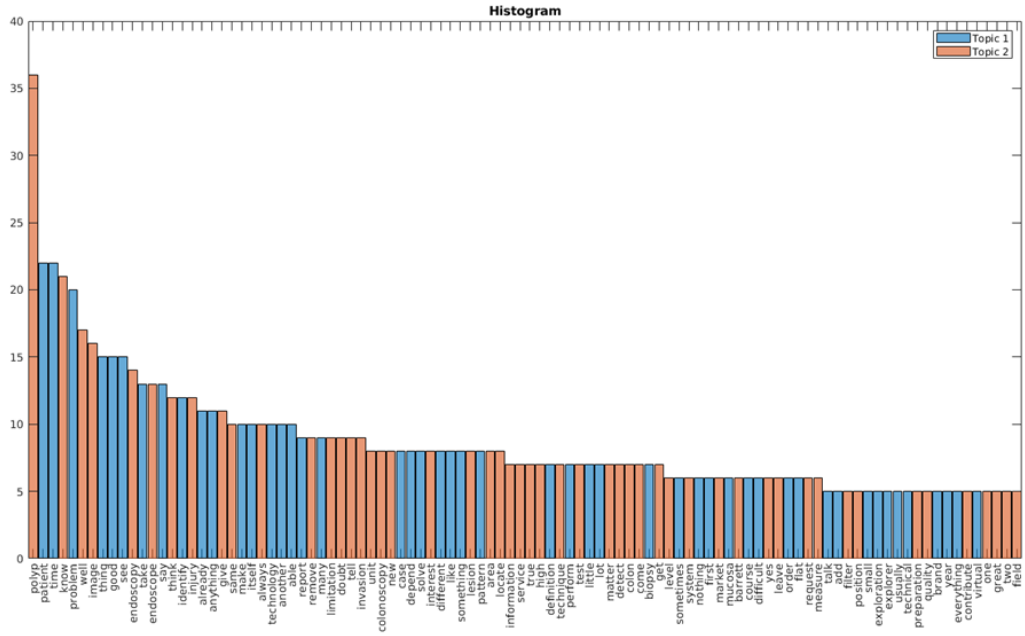


Figure 3. Histogram of the 100 words that appear the most in the interviews

A.2 Unravelling the effect of data augmentation transformations in polyp segmentation

El material suplementario del artículo *Unravelling the effect of data augmentation transformations in polyp segmentation* está compuesto de los siguientes documentos:

A.2.1 - Figuras de todas las transformaciones y rangos para el conjunto de datos CVC-EndoSceneStill

A.2.2 - Figuras de todas las transformaciones y rangos para el conjunto de datos Kvasir-SEG

A.2.1 Figuras de todas las transformaciones y rangos para el conjunto de datos CVC-EndoSceneStill

Unravelling the effect of data augmentation transformations in polyp segmentation
International Journal of Computer Assisted Radiology and Surgery
Luisa F. Sánchez-Peralta*, Artzai Picón, Francisco M. Sánchez-Margallo, J. Blas Pagador
* Jesús Usón Minimally Invasive Surgery Centre. Cáceres (Spain)
<lfsanchez@ccmijesuson.com>

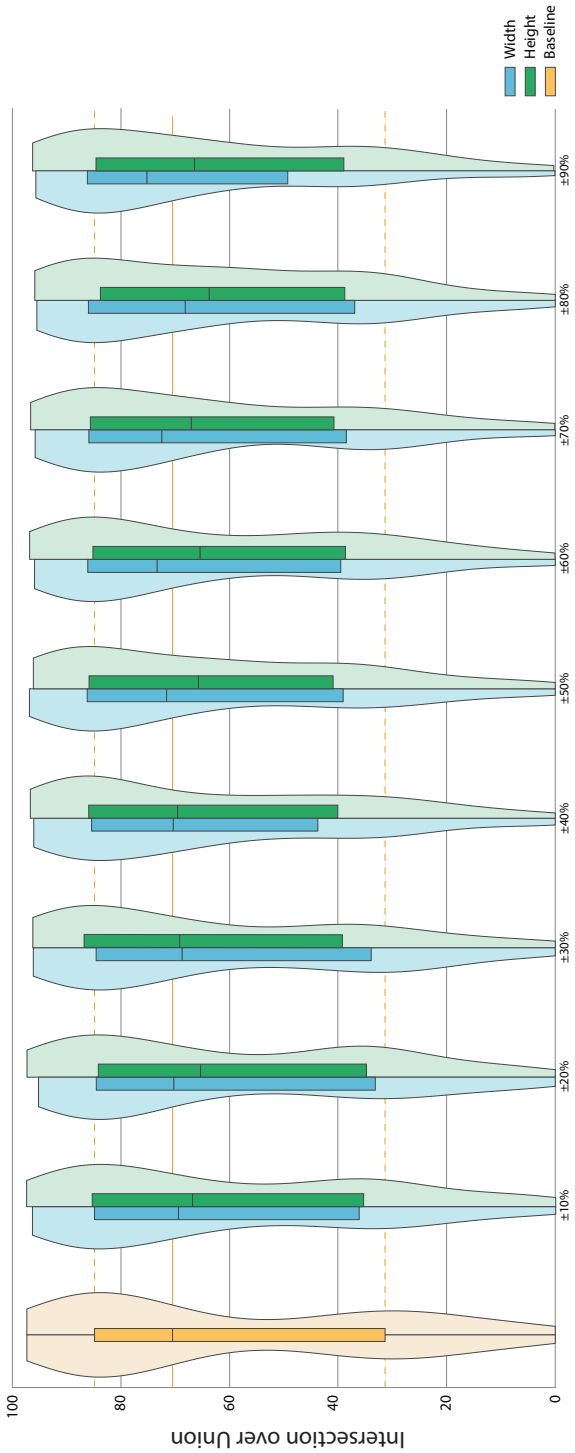


Fig S1.1. Results for width and heigh shifts for CVC-EndoSceneStill. Baseline is included. Its median and quartiles are prolongated on the background for reference.

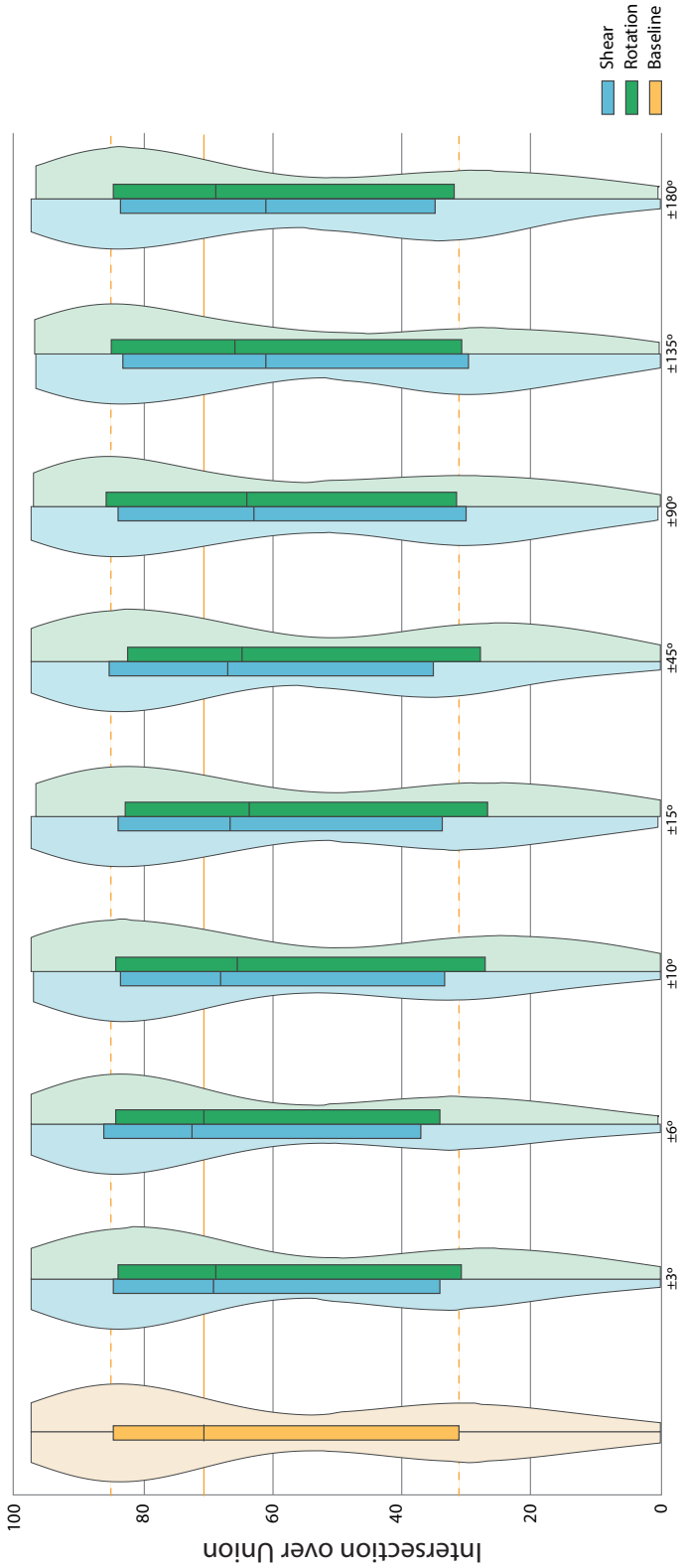


Fig S1.2. Results for rotation and shear for CVC-EndoSceneStill. Baseline is included. Its median and quartiles are prolongated on the background for reference.

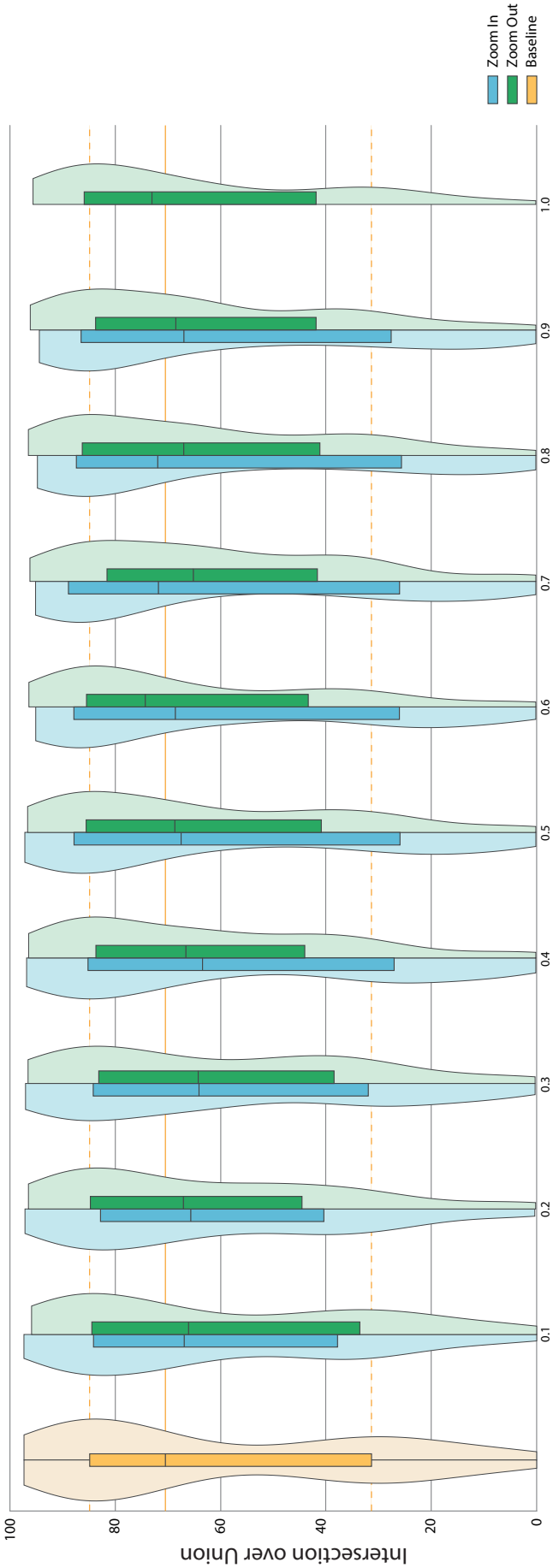


Fig S1.3. Results for zoom in and out for CVC-EndoSceneStill. Baseline is included. Its median and quartiles are prolongated on the background for reference.

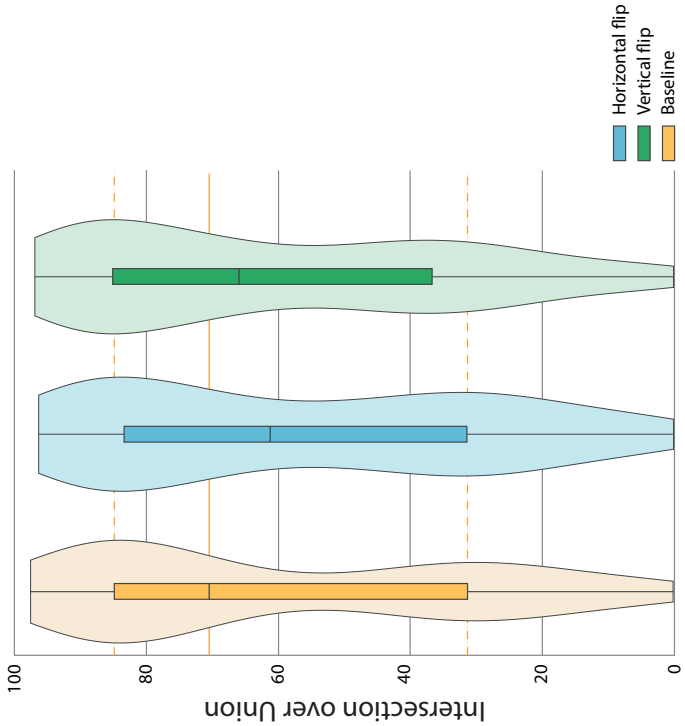


Fig S1.4. Results for horizontal and vertical flips for CVC-EndoSceneStill. Baseline is included. Its median and quartiles are prolongated on the background for reference.

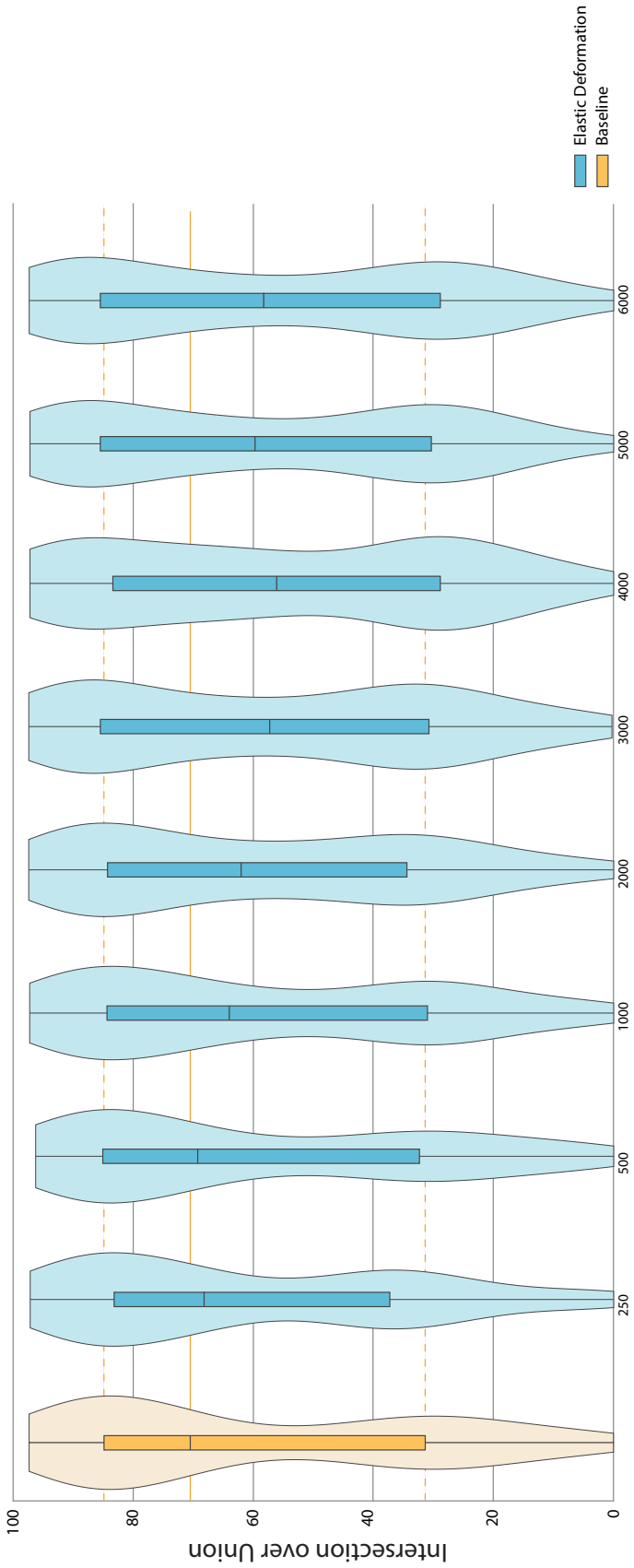


Fig S1.5. Results for elastic deformation for CVC-EndoSceneStill. Baseline is included. Its median and quartiles are prolongated on the background for reference.

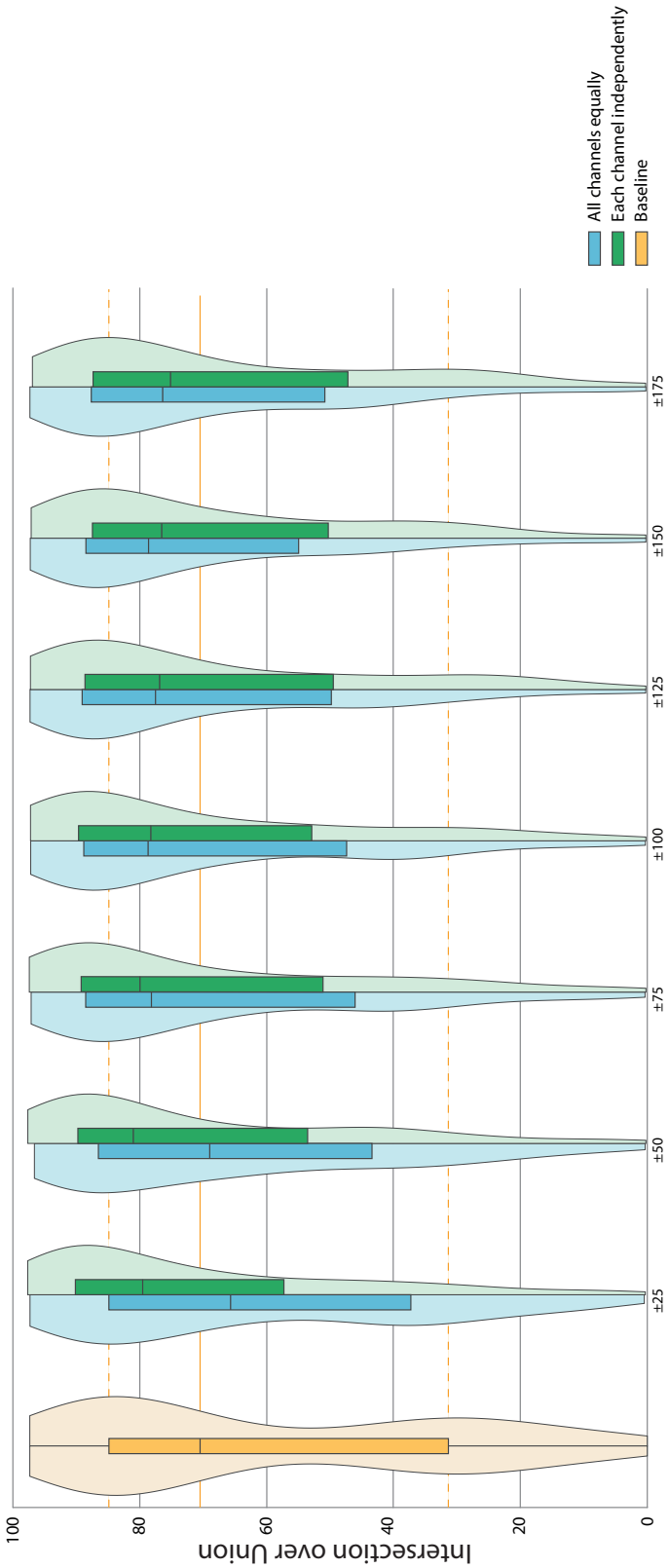


Fig S1.6. Results for changes in brightness for CVC-EndoSceneStill. Baseline is included. Its median and quartiles are prolongated on the background for reference.

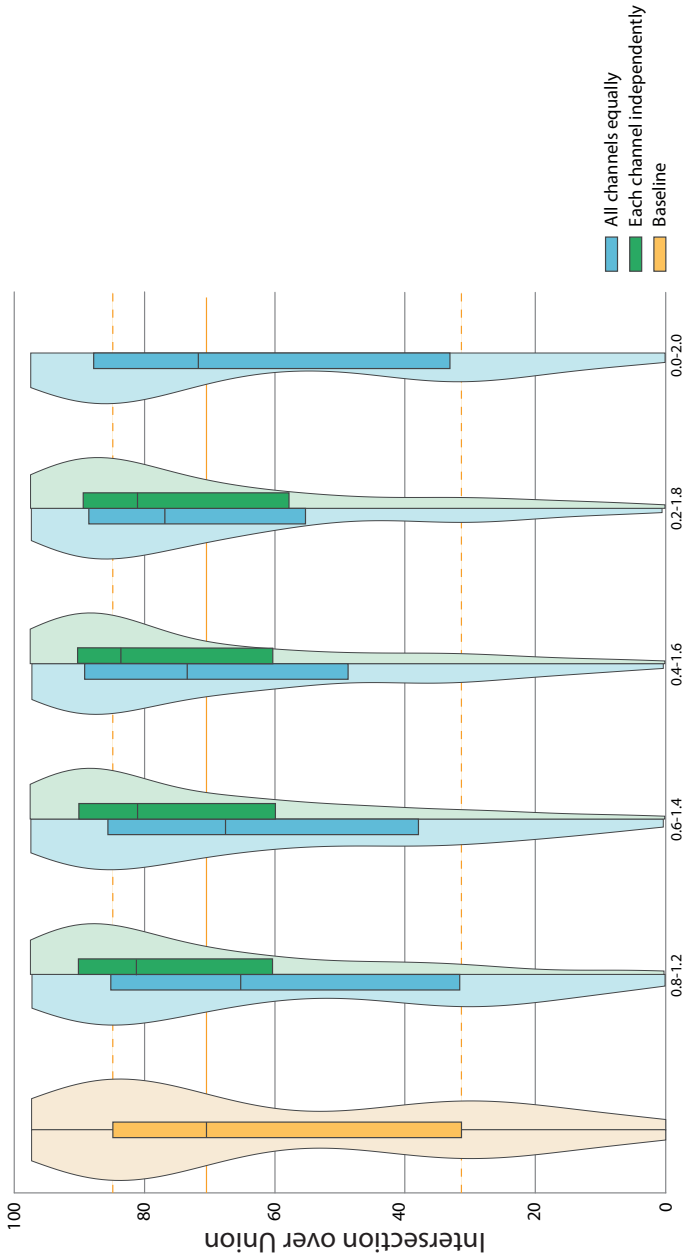


Fig S1.7. Results for changes in contrast for CVC-EndoSceneStill. Baseline is included. Its median and quartiles are prolongated on the background for reference.

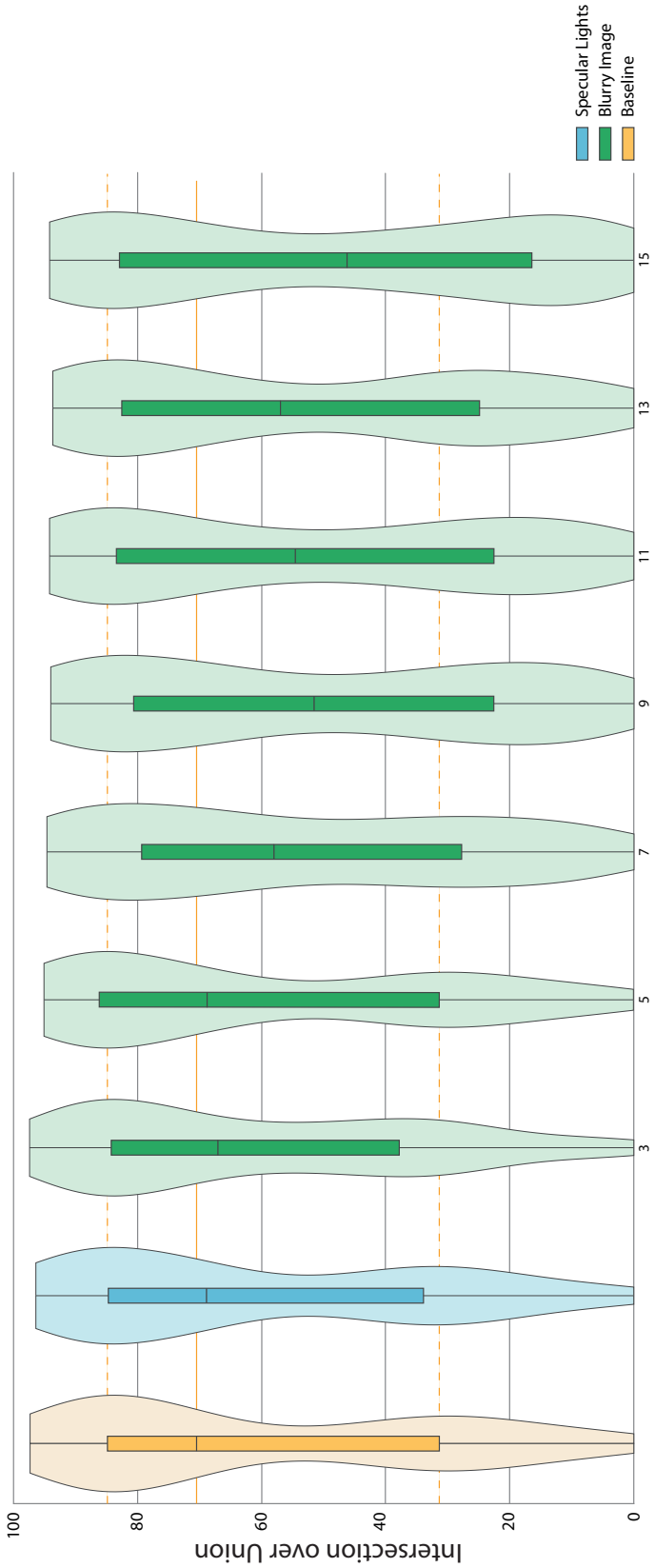


Fig S1.8. Results for inclusion of specular lights and blurry frames (mean filter) for CVC-EndoSceneStill. Baseline is included. Its median and quartiles are prolongated on the background for reference.

A.2.2 Figuras de todas las transformaciones y rangos para el conjunto de datos Kvasir-SEG

Unravelling the effect of data augmentation transformations in polyp segmentation
International Journal of Computer Assisted Radiology and Surgery
Luisa F. Sánchez-Peralta*, Artzai Picón, Francisco M. Sánchez-Margallo, J. Blas Pagador
* Jesús Usón Minimally Invasive Surgery Centre. Cáceres (Spain)
<lfsanchez@ccmijesuson.com>

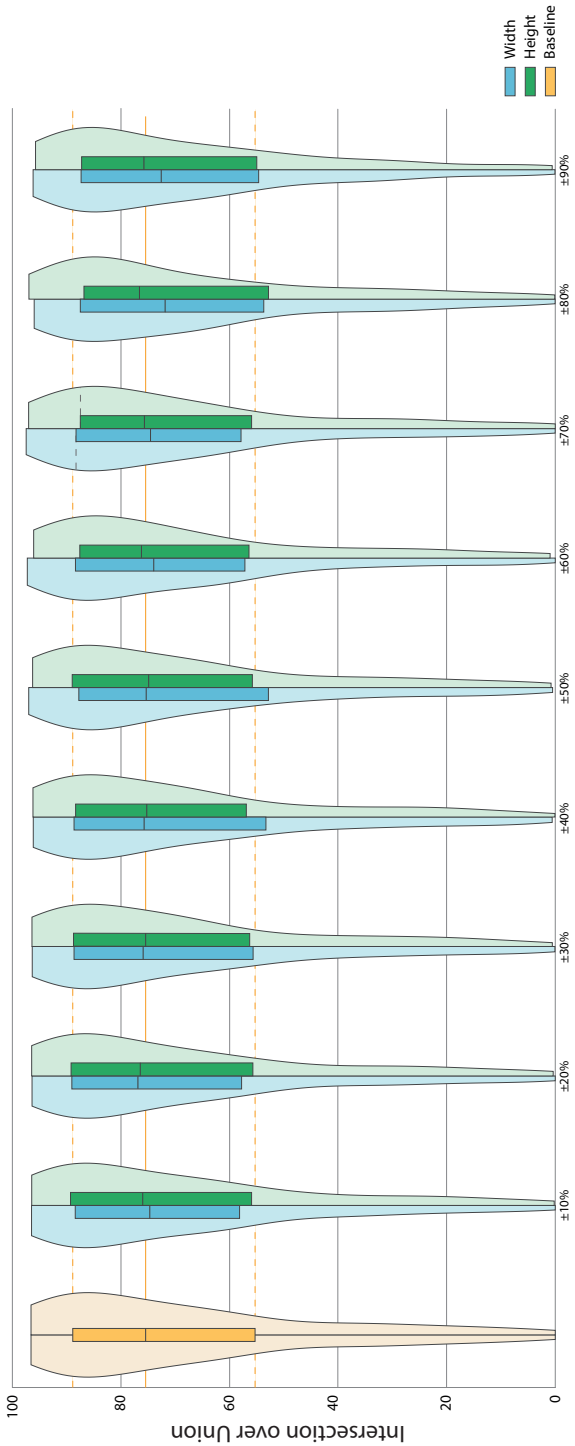


Fig S2.1. Results for width and heigh shifts for Kvasir-SEG. Baseline is included. Its median and quartiles are prolongated on the background for reference.

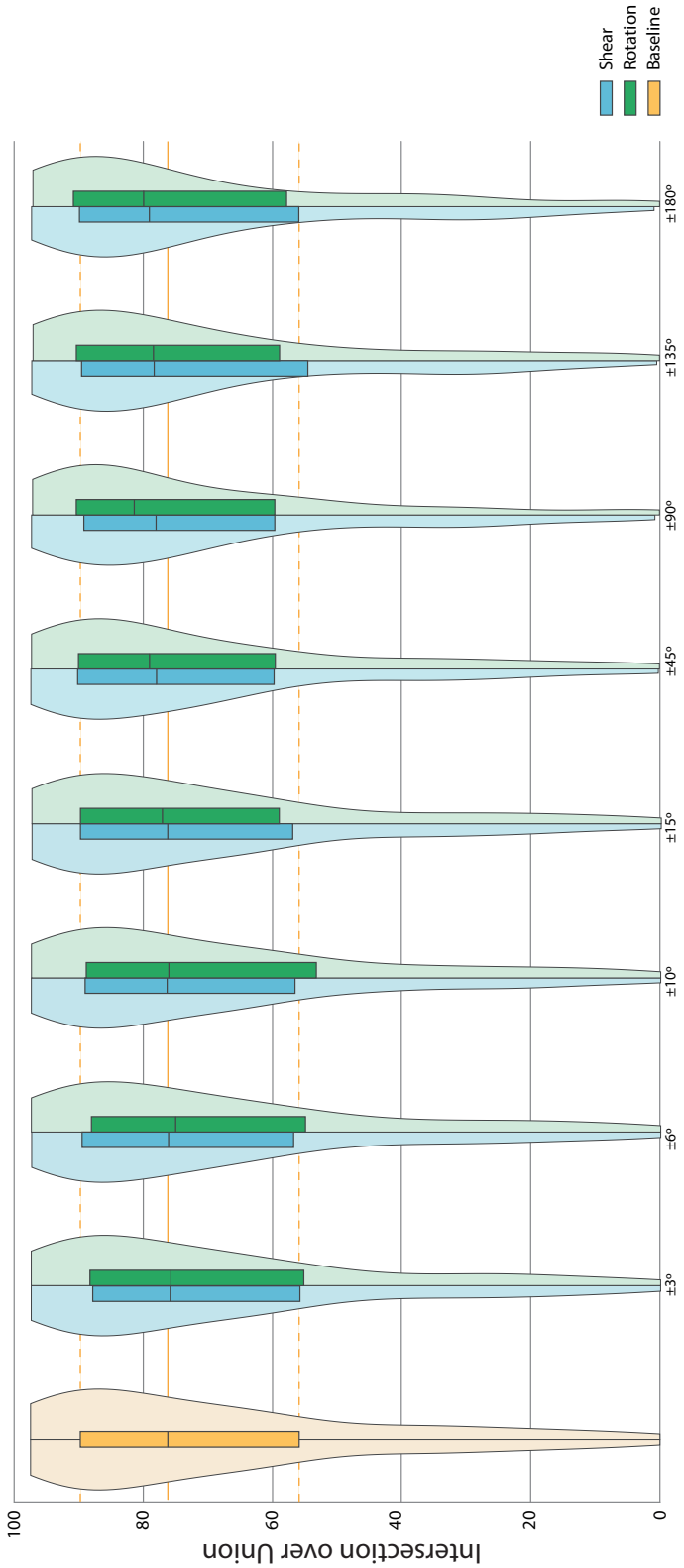


Fig S2.2. Results for rotation and shear for Kvasir-SEG. Baseline is included. Its median and quartiles are prolongated on the background for reference.

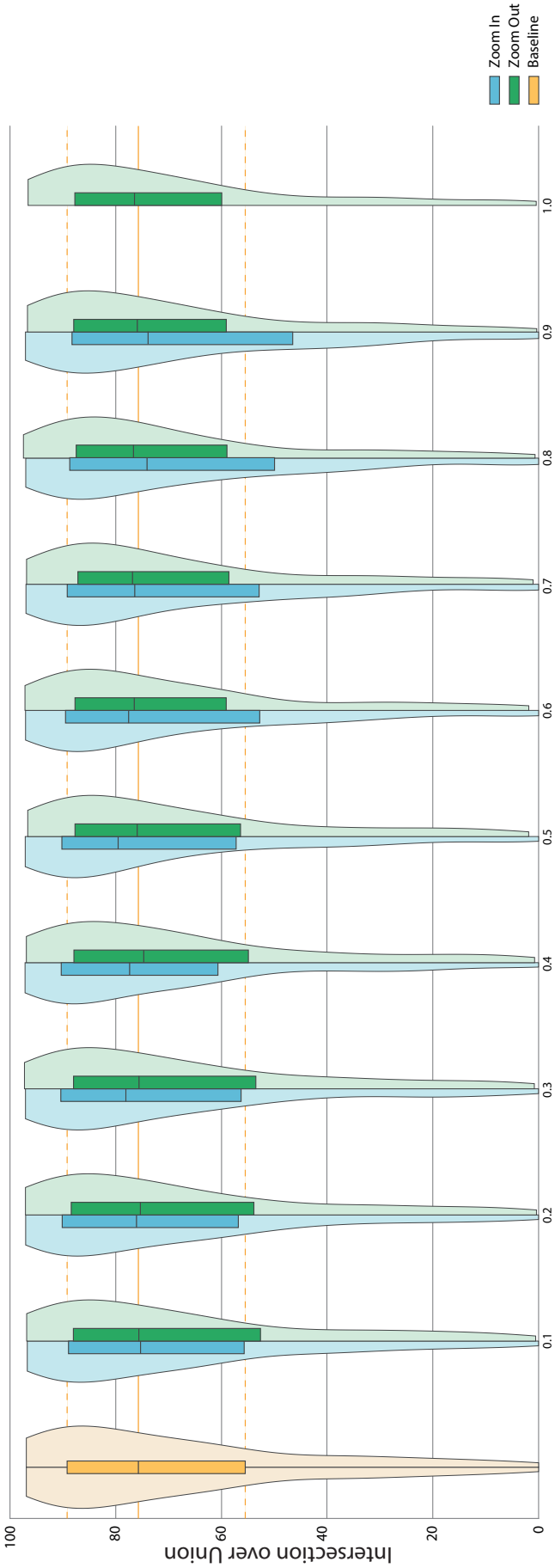


Fig S2.3. Results for zoom in and out for Kvasir-SEG. Baseline is included. Its median and quartiles are prolongated on the background for reference.

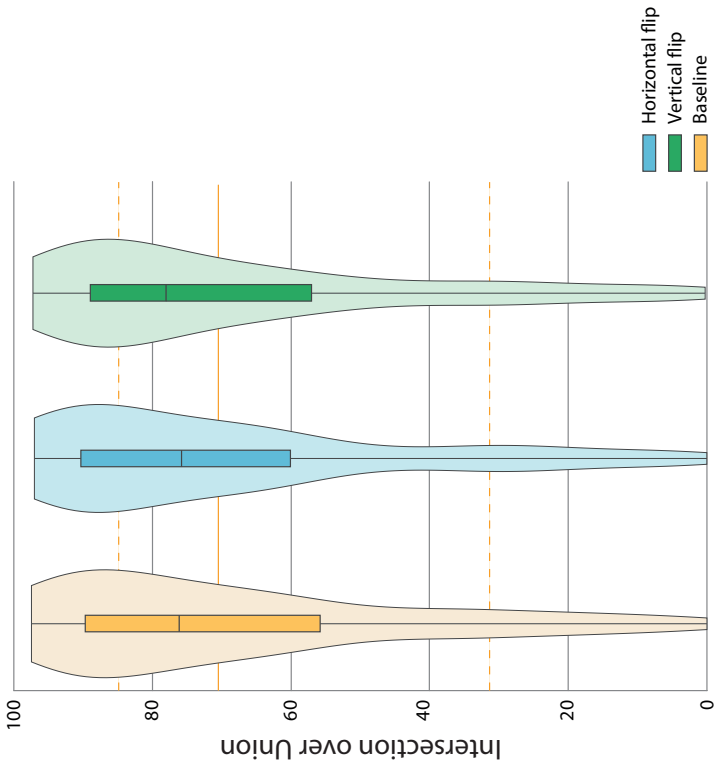


Fig S2.4. Results for horizontal and vertical flips for Kvasir-SEG. Baseline is included. Its median and quartiles are prolongated on the background for reference.

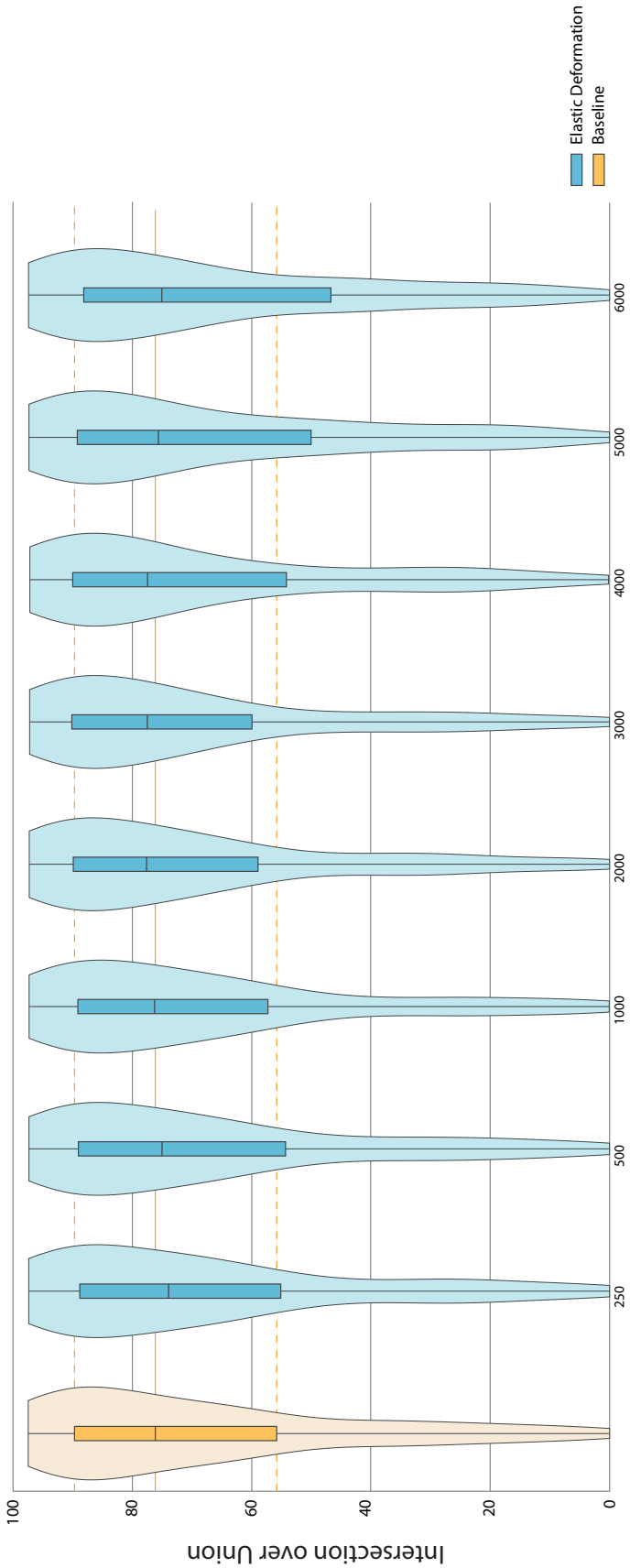


Fig S2.5. Results for elastic deformation for Kvasir-SEG. Baseline is included. Its median and quartiles are prolongated on the background for reference.

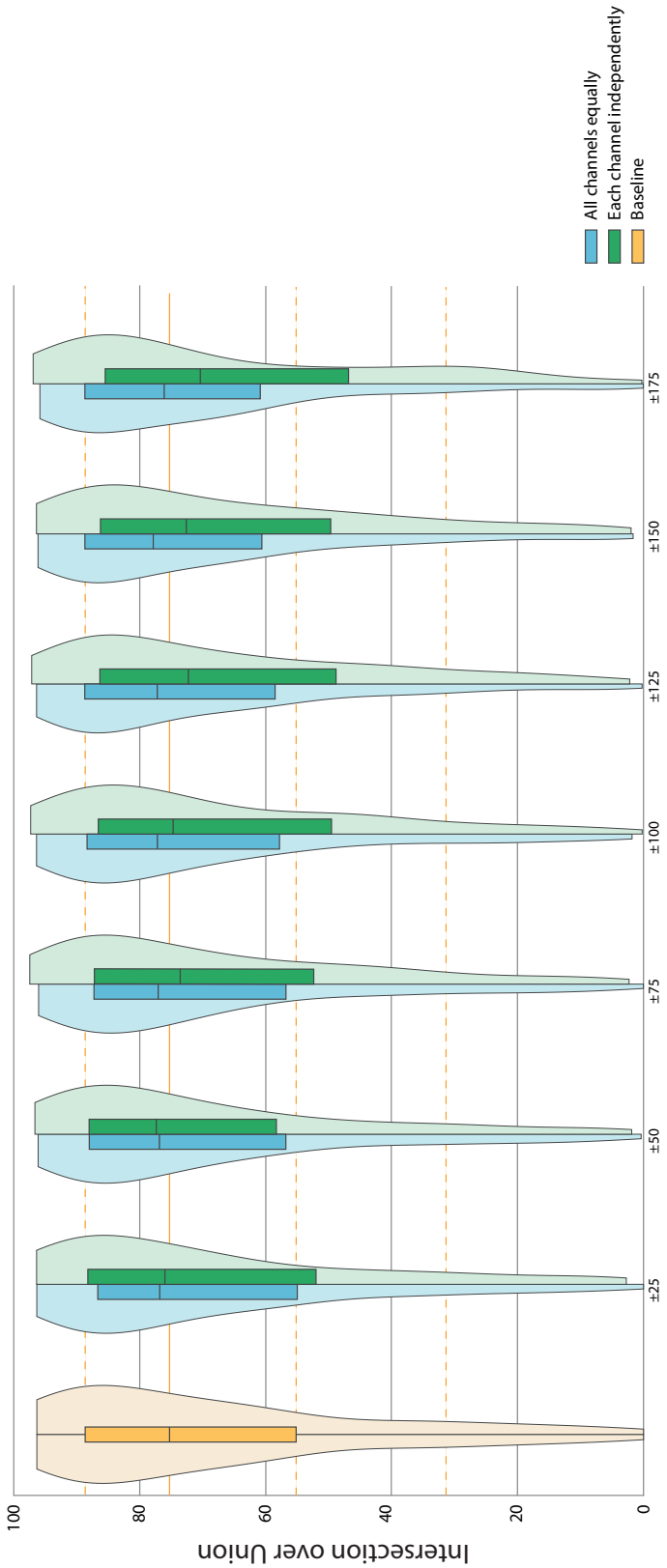


Fig S2.6. Results for changes in brightness for Kvasir-SEG. Baseline is included. Its median and quartiles are prolongated on the background for reference.

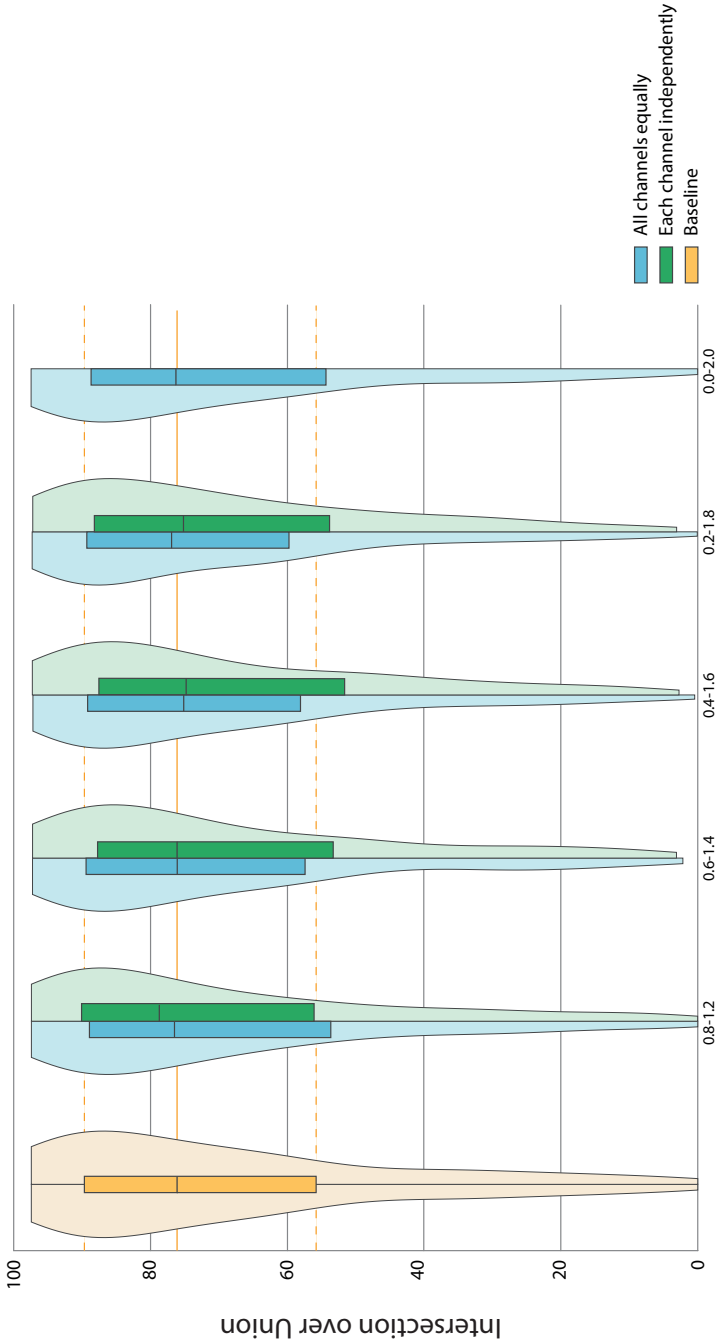


Fig S2.7. Results for changes in contrast for Kvasir-SEG. Baseline is included. Its median and quartiles are prolongated on the background for reference.

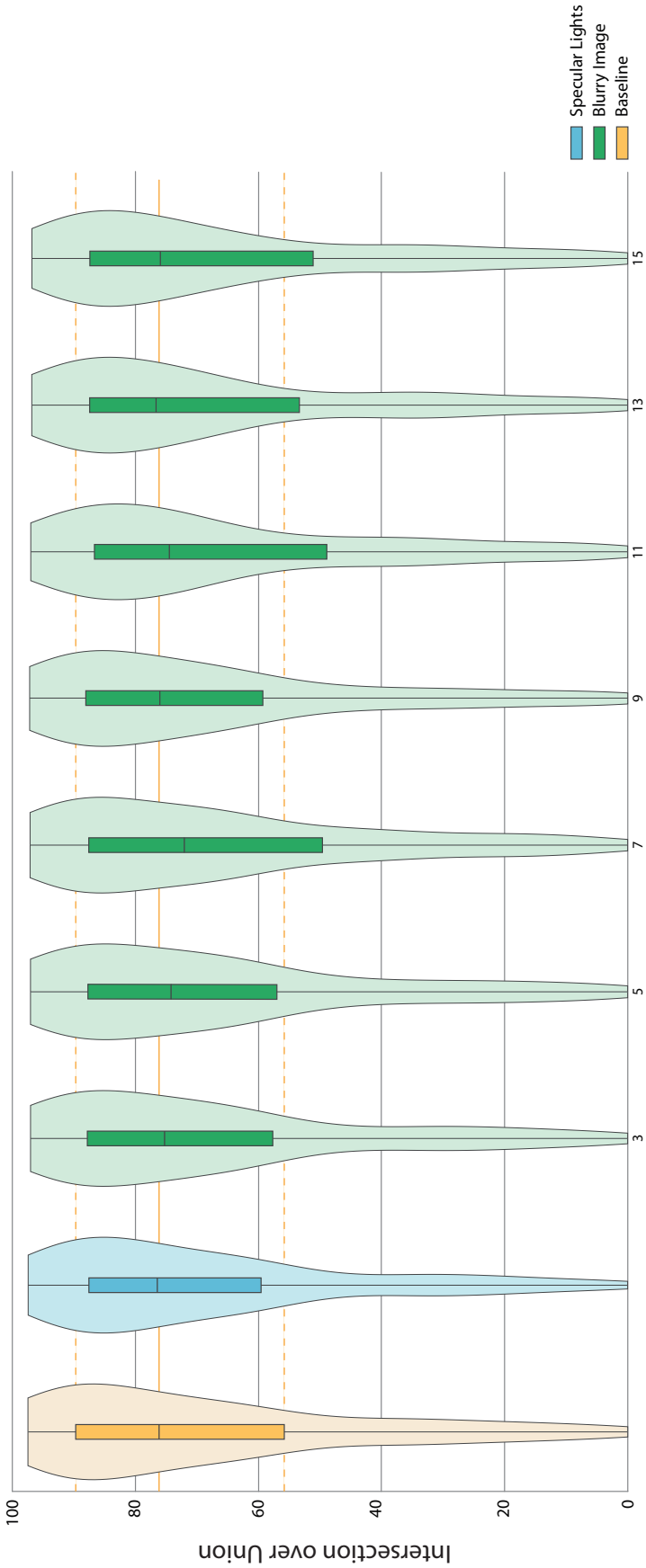


Fig S2.8. Results for inclusion of specular lights and blurry frames (mean filter) for Kvasir-SEG. Baseline is included. Its median and quartiles are prolongated on the background for reference.

A.3 Eigenloss: combined PCA-based loss function for polyp segmentation

El material suplementario del artículo *Eigenloss: combined PCA-based loss function for polyp segmentation* está compuesto de los siguientes documentos:

A.3.1 - Arquitectura U-Net-VGG-16
A.3.2 - Arquitectura U-Net-Densenet121
A.3.3 - Arquitectura LinkNet-VGG-16
A.3.4 - Arquitectura LinkNet-Densenet121
A.3.5 - Diagrama de cajas de la precisión para las diferentes funciones de pérdida y modelos analizados
A.3.6 - Diagrama de cajas de la exhaustividad para las diferentes funciones de pérdida y modelos analizados
A.3.7 - Diagrama de cajas de la especificidad para las diferentes funciones de pérdida y modelos analizados
A.3.8 - Diagrama de cajas del valor F2 para las diferentes funciones de pérdida y modelos analizados
A.3.9 - Diagrama de cajas del índice Dice para las diferentes funciones de pérdida y modelos analizados
A.3.10 - Pérdidas de entrenamiento y validación de U-Net-VGG-16 para todas las funciones de pérdida
A.3.11 - Pérdidas de entrenamiento y validación de U-Net-Densenet121 para todas las funciones de pérdida
A.3.12 - Pérdidas de entrenamiento y validación de LinkNet-VGG-16 para todas las funciones de pérdida
A.3.13 - Pérdidas de entrenamiento y validación de LinkNet-Densenet121 para todas las funciones de pérdida
A.3.14 - Pérdidas, exactitud e índice de Jaccard durante el entrenamiento para los conjuntos de entrenamiento y validación de U-Net-VGG-16
A.3.15 - Pérdidas, exactitud e índice de Jaccard durante el entrenamiento para los conjuntos de entrenamiento y validación de U-Net-Densenet121
A.3.16 - Pérdidas, exactitud e índice de Jaccard durante el entrenamiento para los conjuntos de entrenamiento y validación de LinkNet-VGG-16
A.3.17 - Pérdidas, exactitud e índice de Jaccard durante el entrenamiento para los conjuntos de entrenamiento y validación de LinkNet-Densenet121

A.3.1 Arquitectura U-Net-VGG-16

TableS1. U-Net-VGG-16 architecture. Conv=convolution; ReLU=Rectified Linear Unit

Layer	Output shape	# Parameters	Concatenates
Input	(256×256×3)	-	-
Convolutional Block 1 2×conv (3×3) + max pool	(128×128×64)	38,720	-
Convolutional Block 2 2×conv (3×3) + max pool	(64×64×128)	221,440	-
Convolutional Block 3 3×conv (3×3) + max pool	(32×32×256)	1,475,328	-
Convolutional Block 4 3×conv (3×3) + max pool	(16×16×512)	5,899,776	-
Convolutional Block 5 3×conv (3×3) + max pool	(8×8×512)	7,079,424	-
Center Block 1 conv (3×3) + batch norm + ReLu	(8×8×512)	2,361,344	-
Center Block 2 conv (3×3) + batch norm + ReLu	(8×8×512)	2,361,344	-
Upsampling	(16×16×1024)	-	Convolution in Convolutional Block 5
Decoder stage 0a conv (3×3) + batch norm + ReLu	(16×16×256)	2,360,320	-
Decoder stage 0b conv (3×3) + batch norm + ReLu	(16×16×256)	590,848	-
Upsampling	(32×32×768)	-	Convolution in Convolutional Block 4
Decoder stage 1a conv (3×3) + batch norm + ReLu	(32×32×128)	885,248	-
Decoder stage 1b conv (3×3) + batch norm + ReLu	(32×32×128)	147,968	-
Upsampling	(64×64×384)	-	Convolution in Convolutional Block 3
Decoder stage 2a conv (3×3) + batch norm + ReLu	(64×64×64)	221,440	-
Decoder stage 2b conv (3×3) + batch norm + ReLu	(64×64×64)	37,120	-
Upsampling	(128×128×192)	-	Convolution in Convolutional Block 2
Decoder stage 3a conv (3×3) + batch norm + ReLu	(128×128×32)	55,424	-
Decoder stage 3b conv (3×3) + batch norm + ReLu	(128×128×32)	9,344	-
Upsampling	(256×256×32)	-	-
Decoder stage 4a conv (3×3) + batch norm + ReLu	(256×256×16)	4,672	-
Decoder stage 4b conv (3×3) + batch norm + ReLu	(256×256×16)	2,368	-
Convolution (1×1)	(256×256×1)	145	-
Sigmoid	(256×256×1)	-	-

A.3.2 Arquitectura U-Net-Densenet121

TableS2. U-Net-Densenet121 architecture. Conv=convolution; ReLU=Rectified Linear Unit. Within a dense block, each convolutional block concatenates the previous convolutional layer outputs.

Layer	Output shape	# Parameters	Concatenates
Input	(256×256×3)	-	-
Zero padding	(262×262×3)	-	-
Convolutional block 1 Conv (3×3) + batch norm + ReLu	(128×128×64)	9,664	-
Pool 1 Zero padding + max pool	(64×64×64)	-	-
Dense block 1 $\begin{bmatrix}\text{Batch norm + ReLu + (1 × 1) conv}\\ \text{Batch norm + ReLu + (3 × 3) conv}\end{bmatrix} \times 6$	(64×64×256)	338,304	-
Pool 2 Batch norm + ReLu + conv (1×1) + average pooling	(32×32×128)	33,792	-
Dense block 2 $\begin{bmatrix}\text{Batch norm + ReLu + (1 × 1) conv}\\ \text{Batch norm + ReLu + (3 × 3) conv}\end{bmatrix} \times 12$	(32×32×512)	930,048	-
Pool 3 Batch norm + ReLu + conv (1×1) + average pooling	(16×16×256)	113,120	-
Dense block 3 $\begin{bmatrix}\text{Batch norm + ReLu + (1 × 1) conv}\\ \text{Batch norm + ReLu + (3 × 3) conv}\end{bmatrix} \times 24$	(16×16×1,024)	2,873,856	-
Pool 4 Batch norm + ReLu + conv (1×1) + average pooling	(8×8×512)	528,384	-
Dense block 4 $\begin{bmatrix}\text{Batch norm + ReLu + (1 × 1) conv}\\ \text{Batch norm + ReLu + (3 × 3) conv}\end{bmatrix} \times 16$	(8×8×1024)	2,185,970	-
Batch norm + ReLu	(8×8×1024)	4,096	-
Upsampling	(16×16×1536)	-	Convolution in Pool 4
Decoder stage 0a conv (3×3) + batch norm + ReLu	(16×16×256)	3,539,968	-
Decoder stage 0b conv (3×3) + batch norm + ReLu	(16×16×256)	590,848	-
Upsampling	(32×32×512)	-	Convolution in Pool 3
Decoder stage 1a conv (3×3) + batch norm + ReLu	(32×32×128)	590,336	-
Decoder stage 1b conv (3×3) + batch norm + ReLu	(32×32×128)	147,968	-
Upsampling	(64×64×526)	-	Convolution in Pool 2
Decoder stage 2a conv (3×3) + batch norm + ReLu	(64×64×64)	147,712	-
Decoder stage 2b conv (3×3) + batch norm + ReLu	(64×64×64)	37,120	-
Upsampling	(128×128×128)	-	ReLu in Convolutional Block 1
Decoder stage 3a conv (3×3) + batch norm + ReLu	(128×128×32)	36,992	-
Decoder stage 3b conv (3×3) + batch norm + ReLu	(128×128×32)	9,344	-
Upsampling	(256×256×32)	-	-
Decoder stage 4a conv (3×3) + batch norm + ReLu	(256×256×16)	4,672	-
Decoder stage 4b conv (3×3) + batch norm + ReLu	(256×256×16)	2,368	-
Convolution (1×1)	(256×256×1)	145	-
Sigmoid	(256×256×1)	-	-

A.3.3 Arquitectura LinkNet-VGG-16

TableS3. LinkNet-VGG-16 architecture. Conv=convolution; ReLU=Rectified Linear Unit

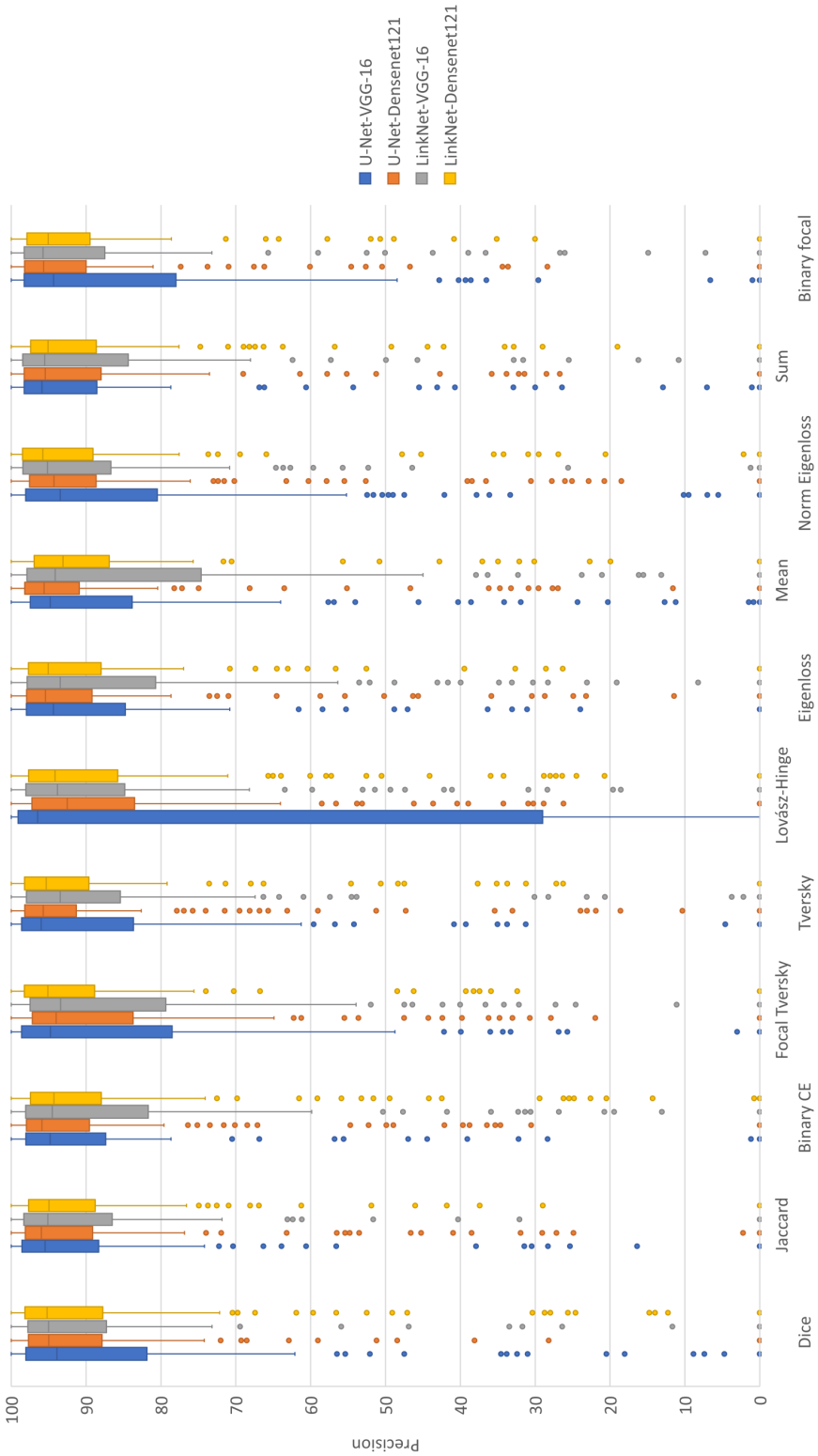
Layer	Output shape	# Parameters	Adds
Input	(256×256×3)	-	-
Convolutional Block 1 2×conv (3×3) + max pool	(128×128×64)	38,720	-
Convolutional Block 2 2×conv (3×3) + max pool	(64×64×128)	221,440	-
Convolutional Block 3 3×conv (3×3) + max pool	(32×32×256)	1,475,328	-
Convolutional Block 4 3×conv (3×3) + max pool	(16×16×512)	5,899,776	-
Convolutional Block 5 3×conv (3×3) + max pool	(8×8×512)	7,079,424	-
Center Block 1 conv (3×3) + batch norm + ReLu	(8×8×512)	2,361,344	-
Center Block 2 conv (3×3) + batch norm + ReLu	(8×8×512)	2,361,344	-
Decoder stage 0a conv (3×3) + batch norm + ReLu + upsampling	(16×16×128)	66,048	-
Decoder stage 0b conv (3×3) + batch norm + ReLu	(16×16×128)	147,968	-
Decoder stage 0c conv (3×3) + batch norm + ReLu	(16×16×512)	67,584	Convolution in Convolutional Block 5
Decoder stage 1a conv (3×3) + batch norm + ReLu + upsampling	(32×32×128)	66,048	-
Decoder stage 1b conv (3×3) + batch norm + ReLu	(32×32×128)	147,968	-
Decoder stage 1c conv (3×3) + batch norm + ReLu	(32×32×512)	67,584	Convolution in Convolutional Block 4
Decoder stage 2a conv (3×3) + batch norm + ReLu + upsampling	(64×64×128)	66,048	-
Decoder stage 2b conv (3×3) + batch norm + ReLu	(64×64×128)	147,968	-
Decoder stage 2c conv (3×3) + batch norm + ReLu	(64×64×256)	33,792	Convolution in Convolutional Block 3
Decoder stage 3a conv (3×3) + batch norm + ReLu + upsampling	(128×128×64)	16,640	-
Decoder stage 3b conv (3×3) + batch norm + ReLu	(128×128×64)	37,120	-
Decoder stage 3c conv (3×3) + batch norm + ReLu	(128×128×128)	8,704	Convolution in Convolutional Block 2
Decoder stage 4a conv (3×3) + batch norm + ReLu + upsampling	(256×256×32)	4,224	-
Decoder stage 4b conv (3×3) + batch norm + ReLu	(256×256×32)	9,344	-
Decoder stage 4c conv (3×3) + batch norm + ReLu	(256×256×16)	576	-
Convolution (1×1)	(256×256×1)	145	-
Sigmoid	(256×256×1)	-	-

A.3.4 Arquitectura LinkNet-Densenet121

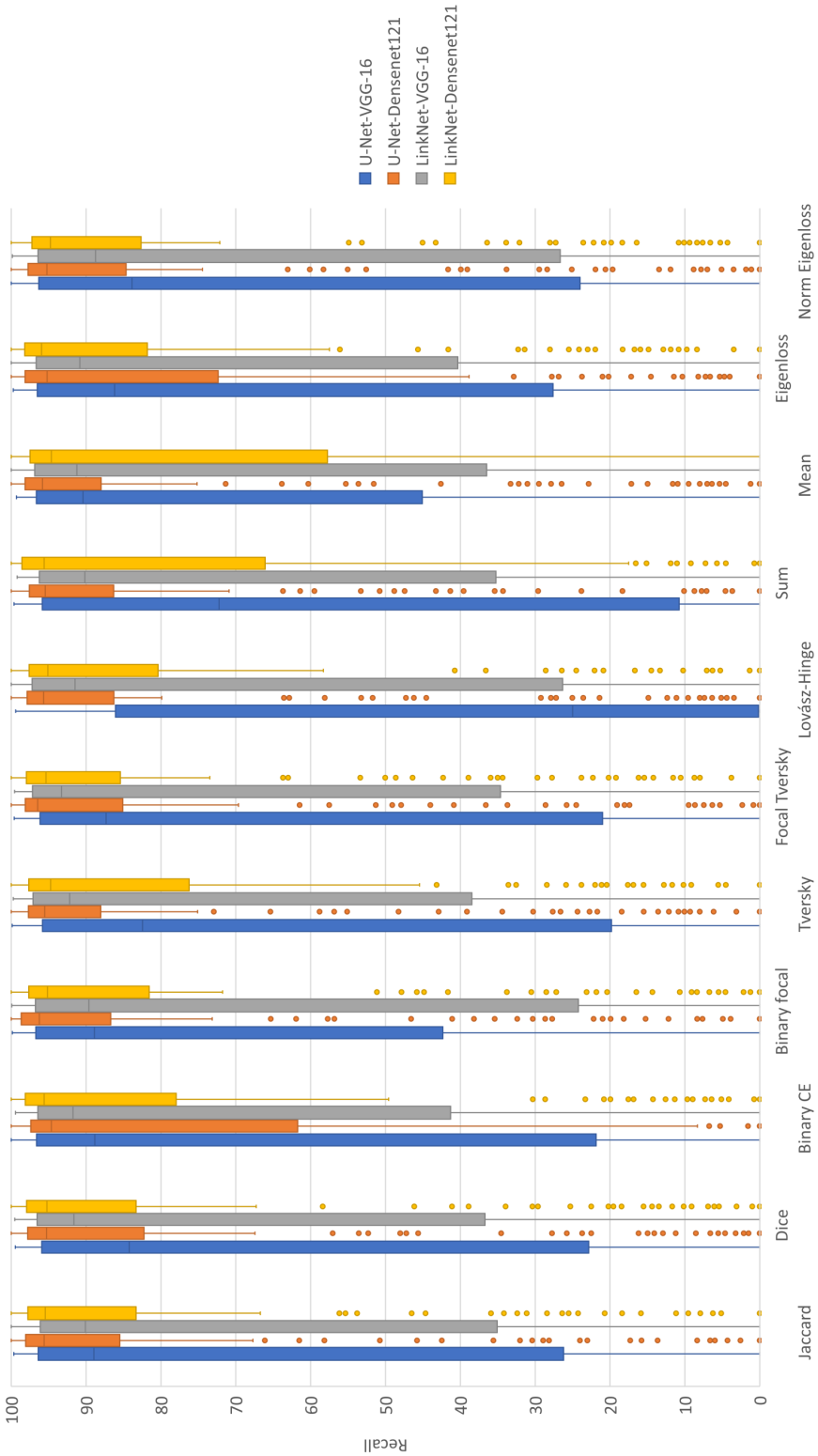
TableS4. Linknet-Densenet121 architecture. Conv=convolution; ReLU=Rectified Linear Unit. Within a dense block, each convolutional block concatenates the previous convolutional layer outputs.

Layer	Output shape	# Parameters	Adds
Input	(256×256×3)	-	-
Zero padding	(262×262×3)	-	-
Conv (3×3) + batch norm + ReLu	(128×128×64)	9,664	-
Pool 1 Zero padding + max pool 1	(64×64×64)	-	-
Dense block 1 $\begin{bmatrix}\text{Batch norm + ReLu + (1 × 1) conv}\\ \text{Batch norm + ReLu + (3 × 3) conv}\end{bmatrix} \times 6$	(64×64×256)	338,304	-
Pool 2 Batch norm + ReLu + conv (1×1) + average pooling	(32×32×128)	33,792	-
Dense block 2 $\begin{bmatrix}\text{Batch norm + ReLu + (1 × 1) conv}\\ \text{Batch norm + ReLu + (3 × 3) conv}\end{bmatrix} \times 12$	(32×32×512)	930,048	-
Pool 3 Batch norm + ReLu + conv (1×1) + average pooling	(16×16×256)	113,120	-
Dense block 3 $\begin{bmatrix}\text{Batch norm + ReLu + (1 × 1) conv}\\ \text{Batch norm + ReLu + (3 × 3) conv}\end{bmatrix} \times 24$	(16×16×1,024)	2,873,856	
Pool 4 Batch norm + ReLu + conv (1×1) + average pooling	(8×8×512)	528,384	
Dense block 4 $\begin{bmatrix}\text{Batch norm + ReLu + (1 × 1) conv}\\ \text{Batch norm + ReLu + (3 × 3) conv}\end{bmatrix} \times 16$	(8×8×1024)	2,185,970	
Batch norm + ReLu	(8×8×1024)	4,096	
Decoder stage 0a conv (3×3) + batch norm + ReLu + upsampling	(16×16×256)	263,168	-
Decoder stage 0b conv (3×3) + batch norm + ReLu	(16×16×128)	590,848	-
Decoder stage 0c conv (3×3) + batch norm + ReLu	(16×16×512)	133,120	Convolution in Pool 4
Decoder stage 1a conv (3×3) + batch norm + ReLu + upsampling	(32×32×128)	66,048	-
Decoder stage 1b conv (3×3) + batch norm + ReLu	(32×32×128)	147,968	
Decoder stage 1c conv (3×3) + batch norm + ReLu	(32×32×256)	33,792	Convolution in Pool 3
Decoder stage 2a conv (3×3) + batch norm + ReLu + upsampling	(64×64×64)	16,640	
Decoder stage 2b conv (3×3) + batch norm + ReLu	(64×64×64)	37,120	
Decoder stage 2c conv (3×3) + batch norm + ReLu	(64×64×128)	8,704	Convolution in Pool 2
Decoder stage 3a conv (3×3) + batch norm + ReLu + upsampling	(128×128×32)	4,224	
Decoder stage 3b conv (3×3) + batch norm + ReLu	(128×128×32)	9,344	
Decoder stage 3c conv (3×3) + batch norm + ReLu	(128×128×64)	2,304	ReLu in Convolutional Block 1
Decoder stage 4a conv (3×3) + batch norm + ReLu + upsampling	(256×256×16)	1,088	-
Decoder stage 4b conv (3×3) + batch norm + ReLu	(256×256×16)	2,368	-
Decoder stage 4c conv (3×3) + batch norm + ReLu	(256×256×16)	320	
Convolution (1×1)	(256×256×1)	145	-
Sigmoid	(256×256×1)	-	-

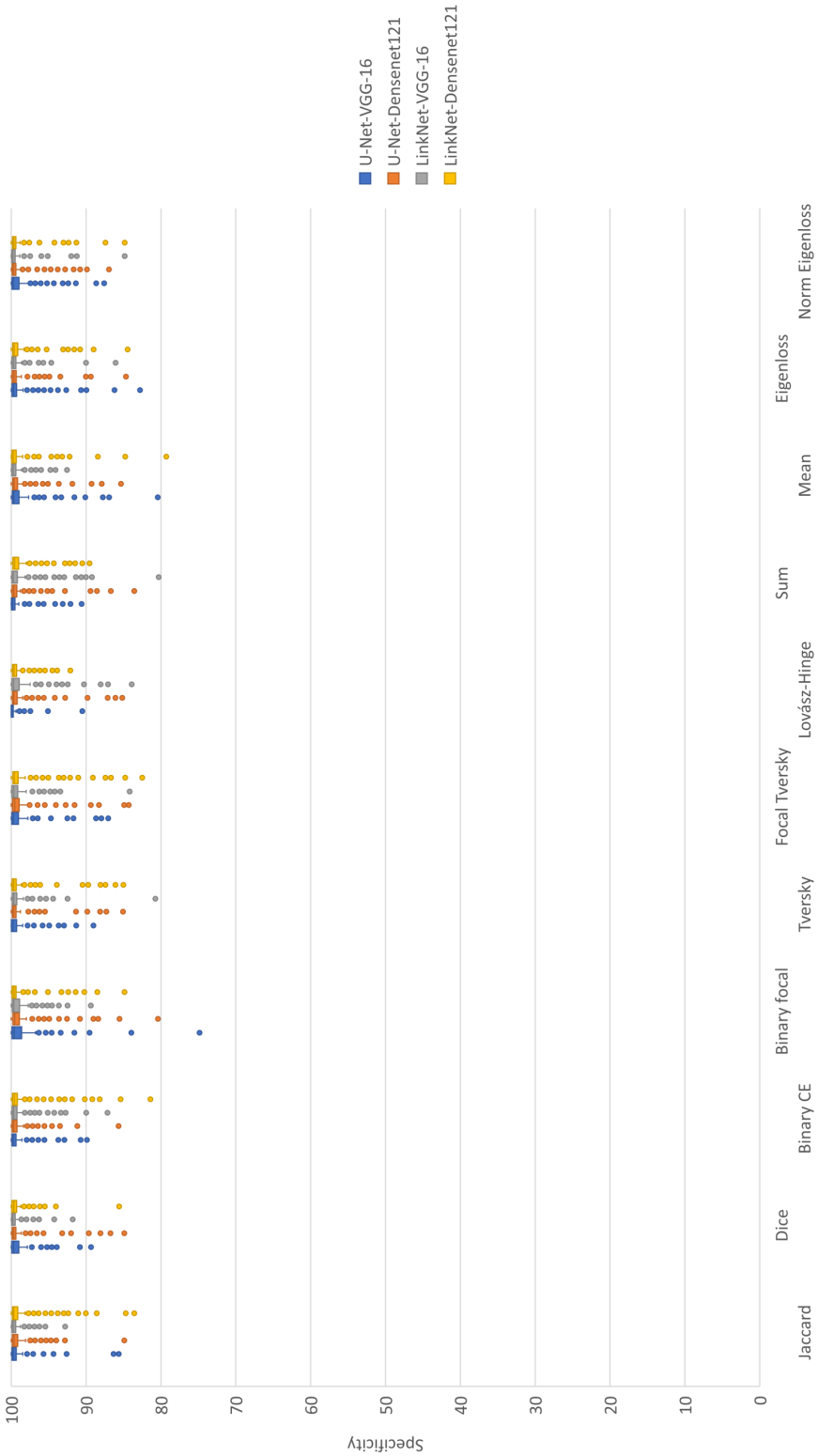
A.3.5 Diagrama de cajas de la precisión para las diferentes funciones de pérdida y modelos analizados



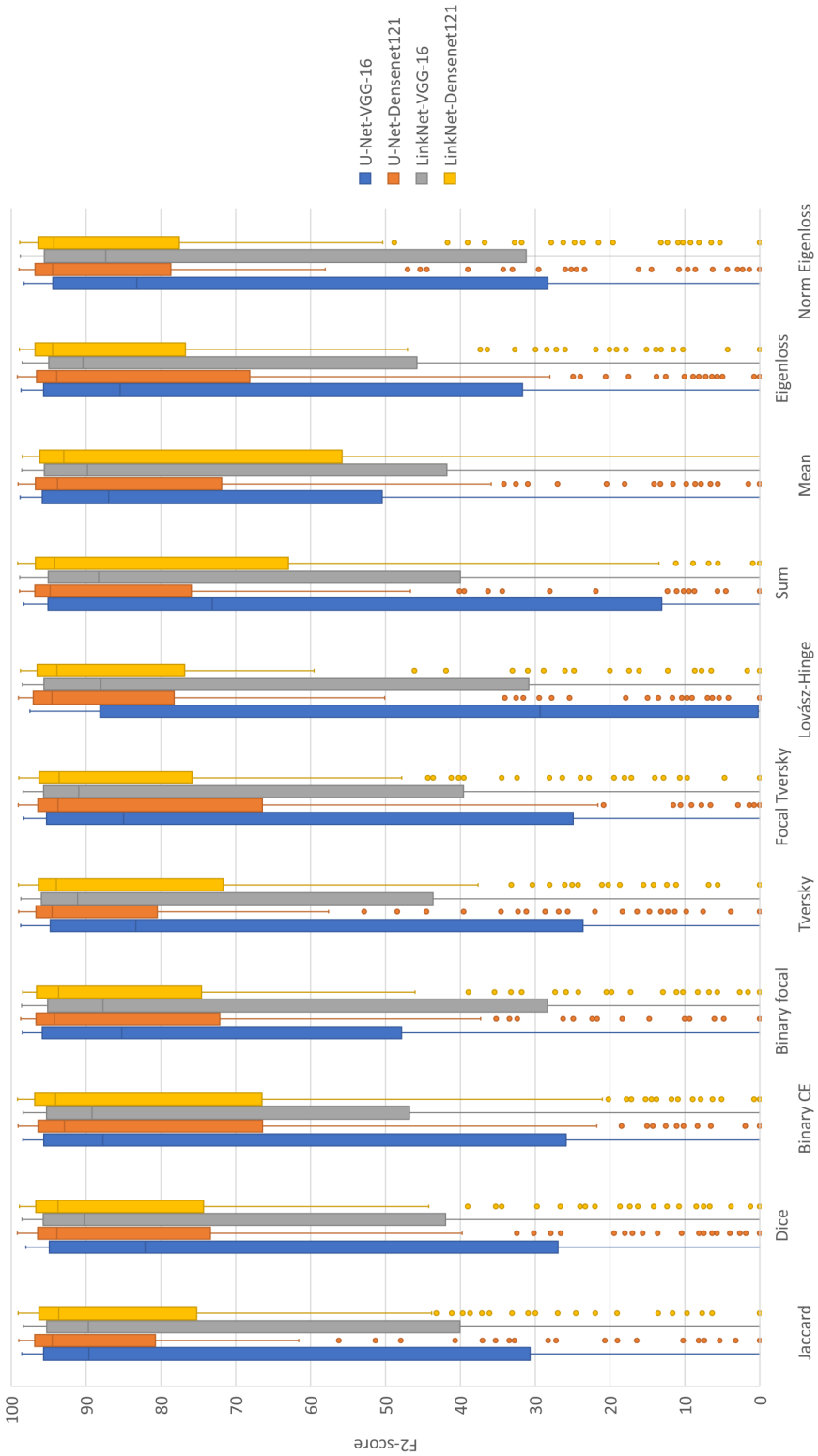
A.3.6 Diagrama de cajas de la exhaustividad para las diferentes funciones de pérdida y modelos analizados



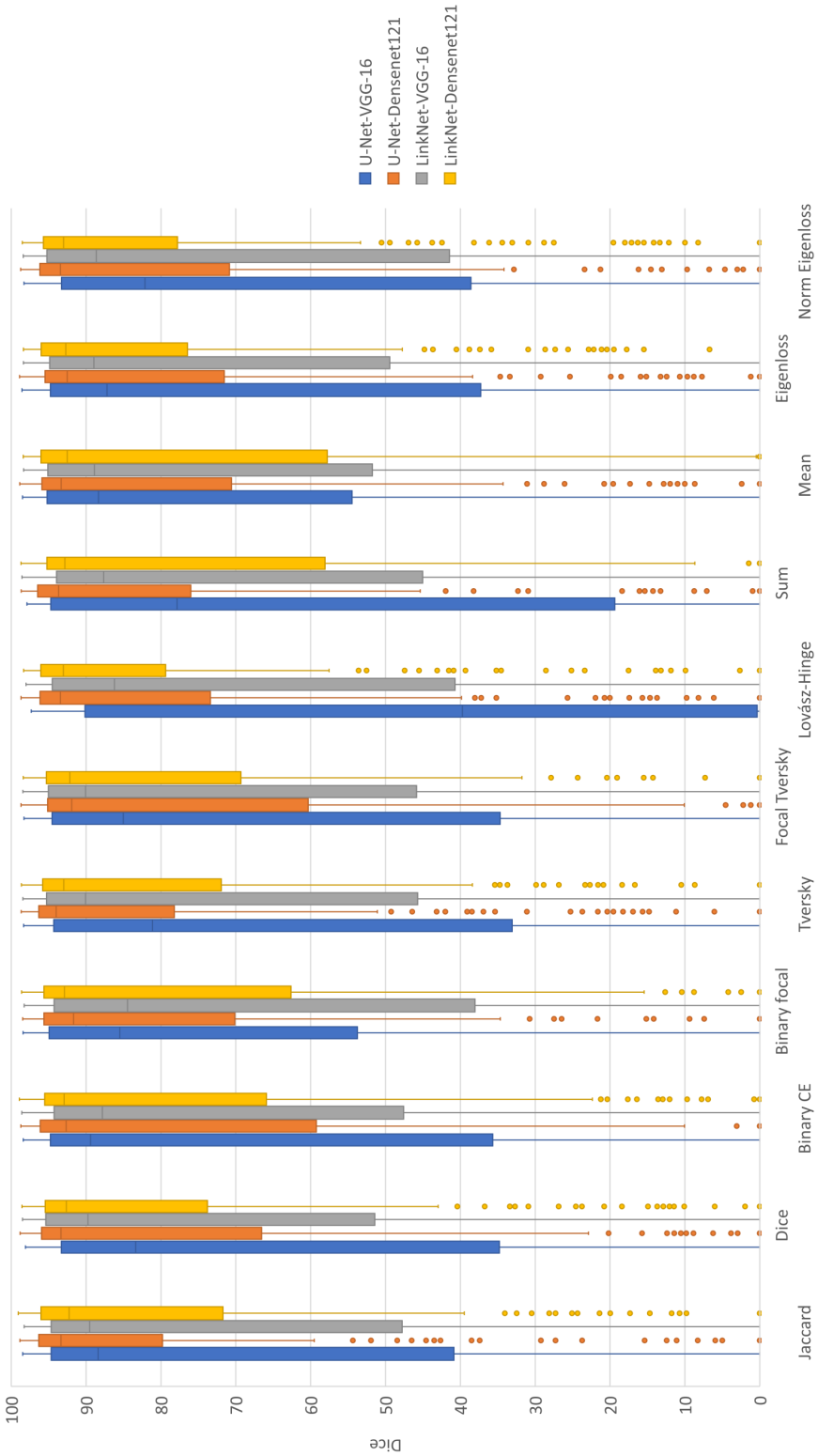
A.3.7 Diagrama de cajas de la especificidad para las diferentes funciones de pérdida y modelos analizados



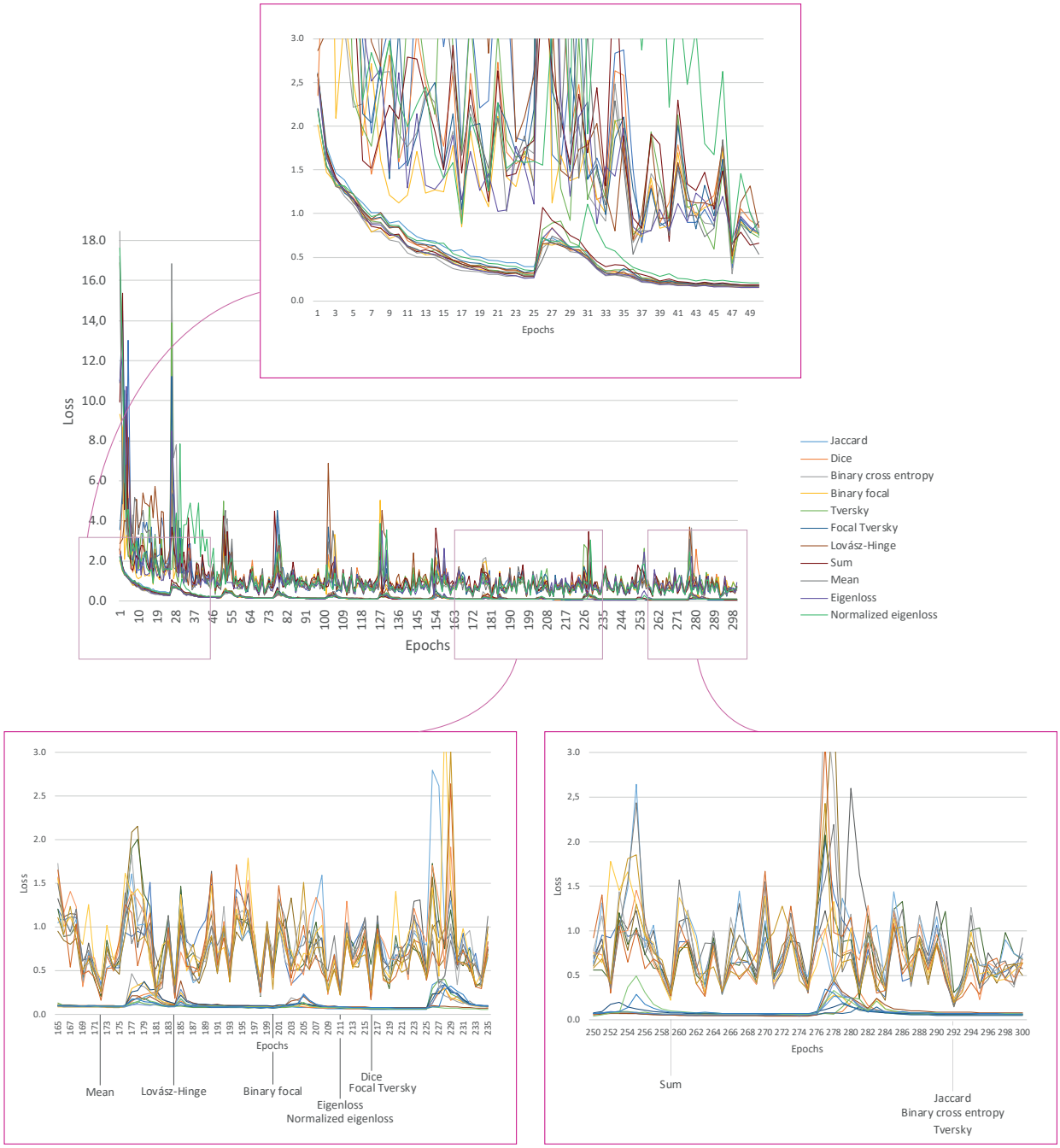
A.3.8 Diagrama de cajas del valor F2 para las diferentes funciones de pérdida y modelos analizados



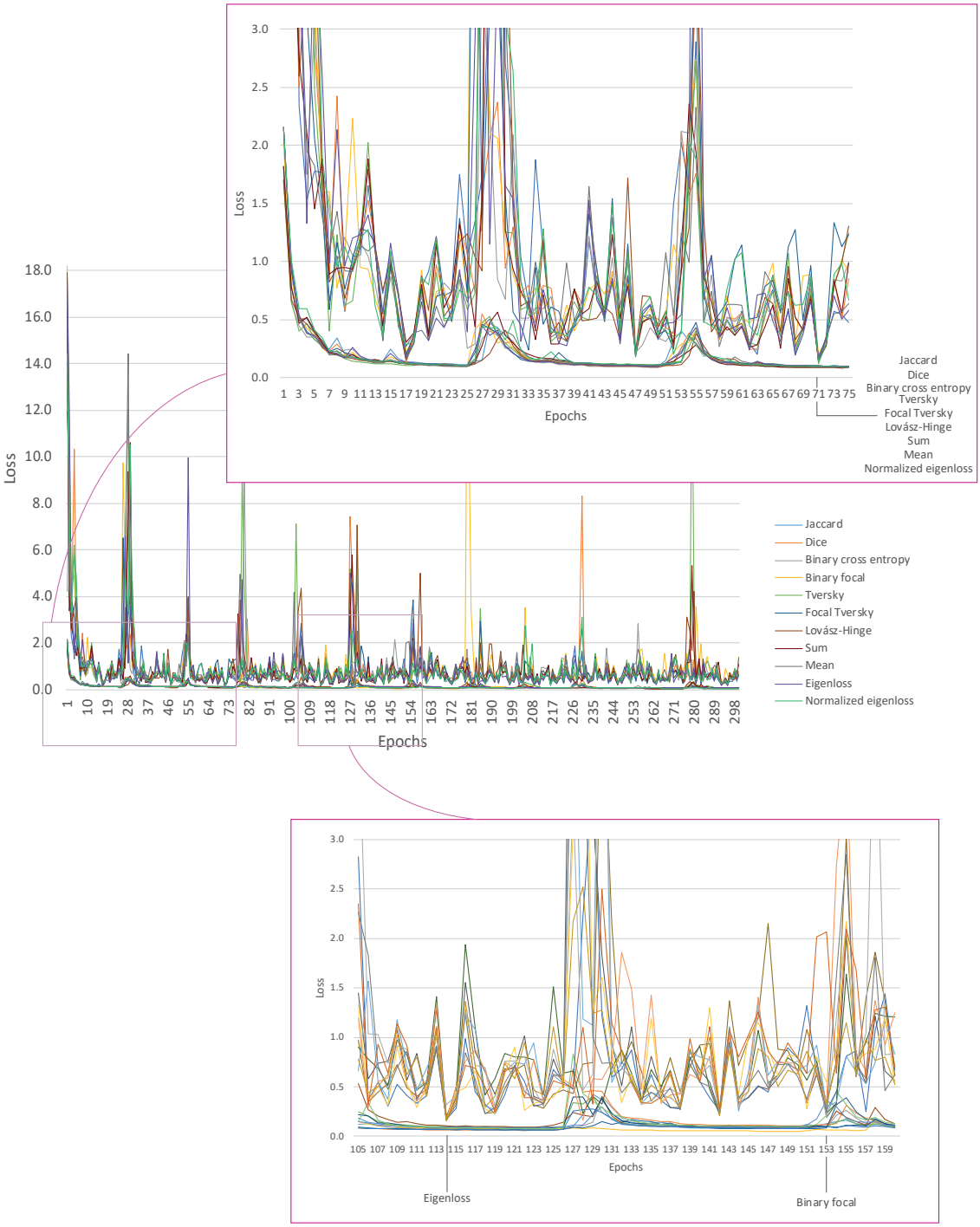
A.3.9 Diagrama de cajas del índice Dice para las diferentes funciones de pérdida y modelos analizados



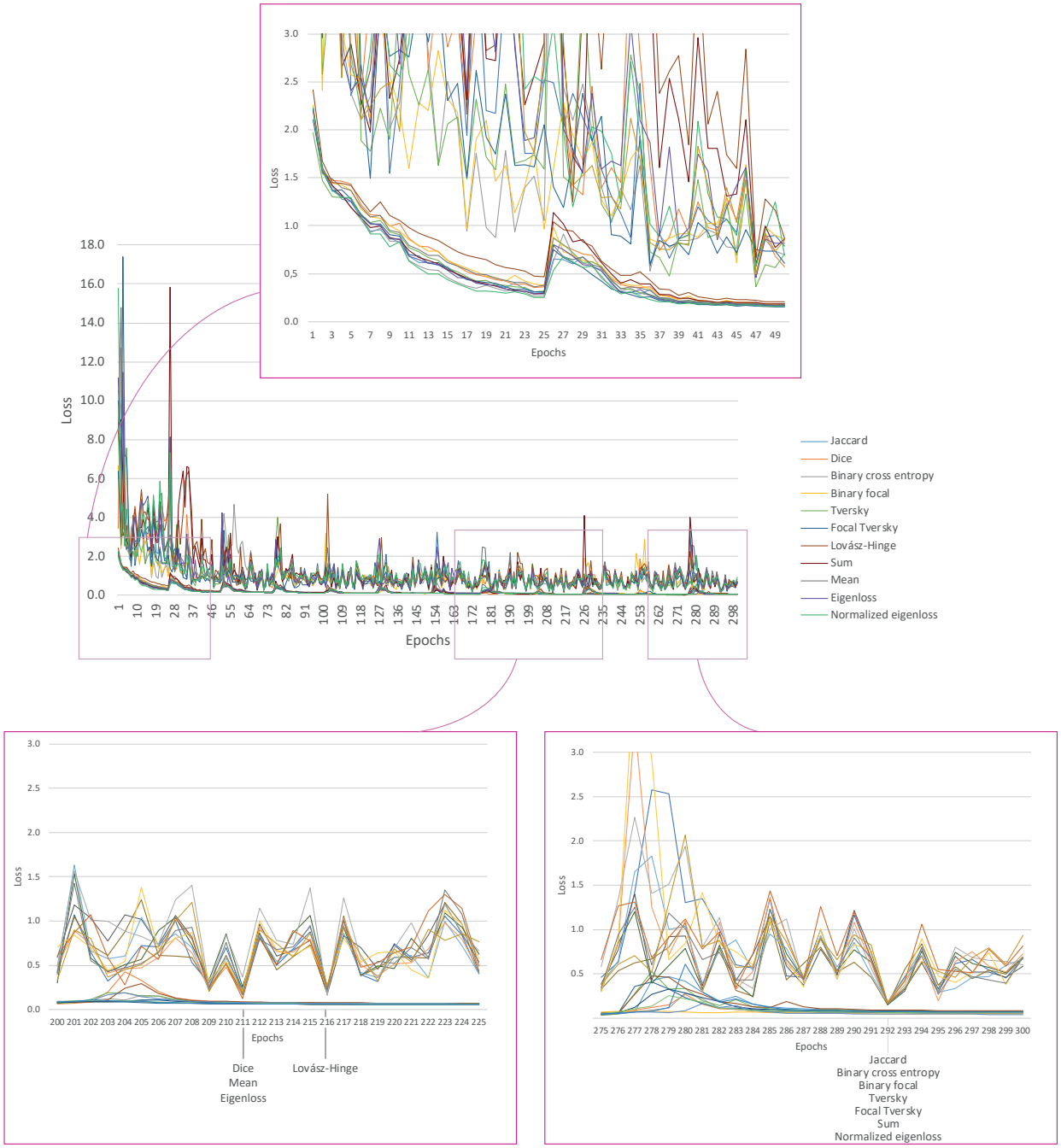
A.3.10 Pérdidas de entrenamiento y validación de U-Net-VGG-16 para todas las funciones de pérdida



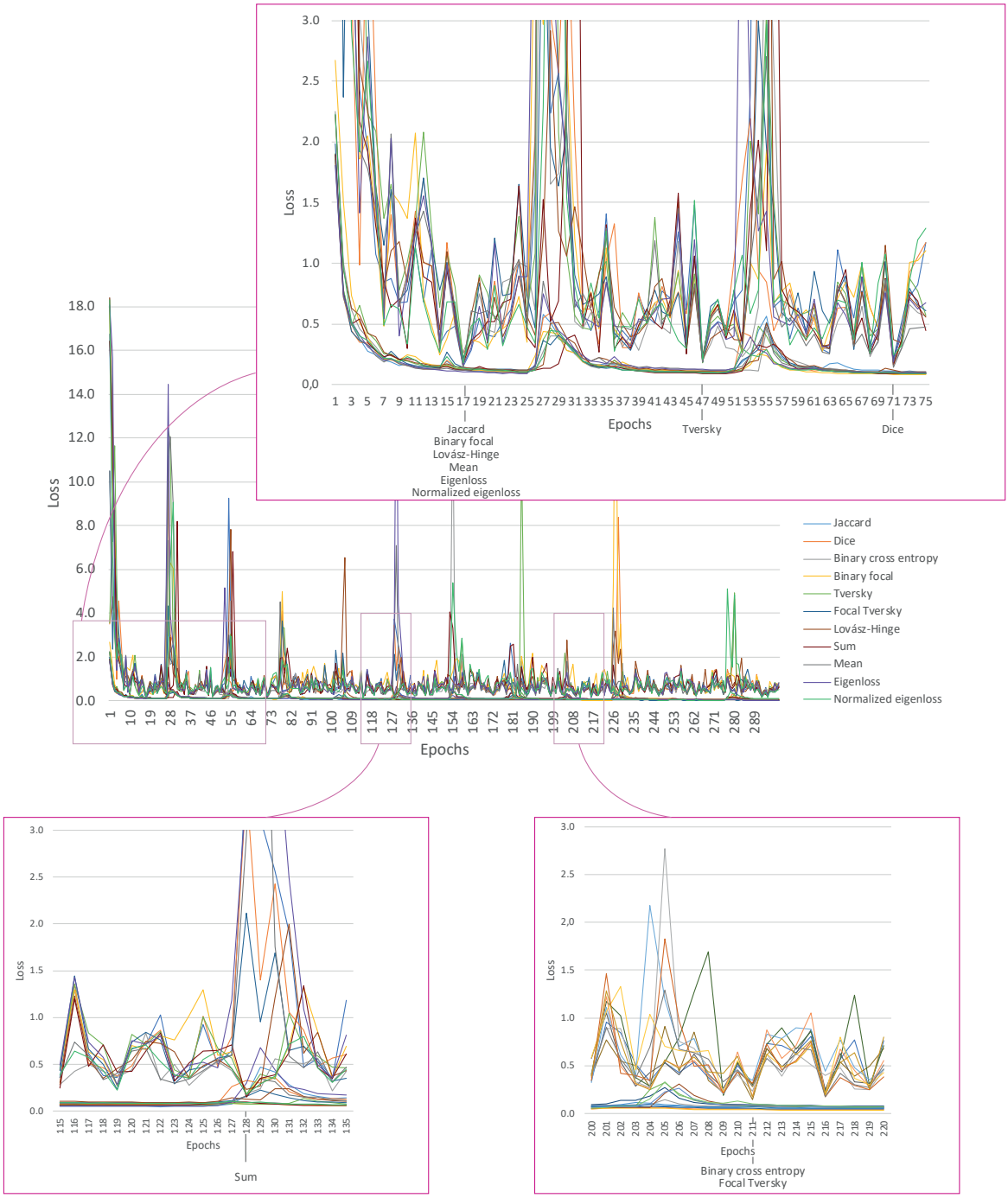
A.3.11 Pérdidas de entrenamiento y validación de U-Net-Densenet121 para todas las funciones de pérdida



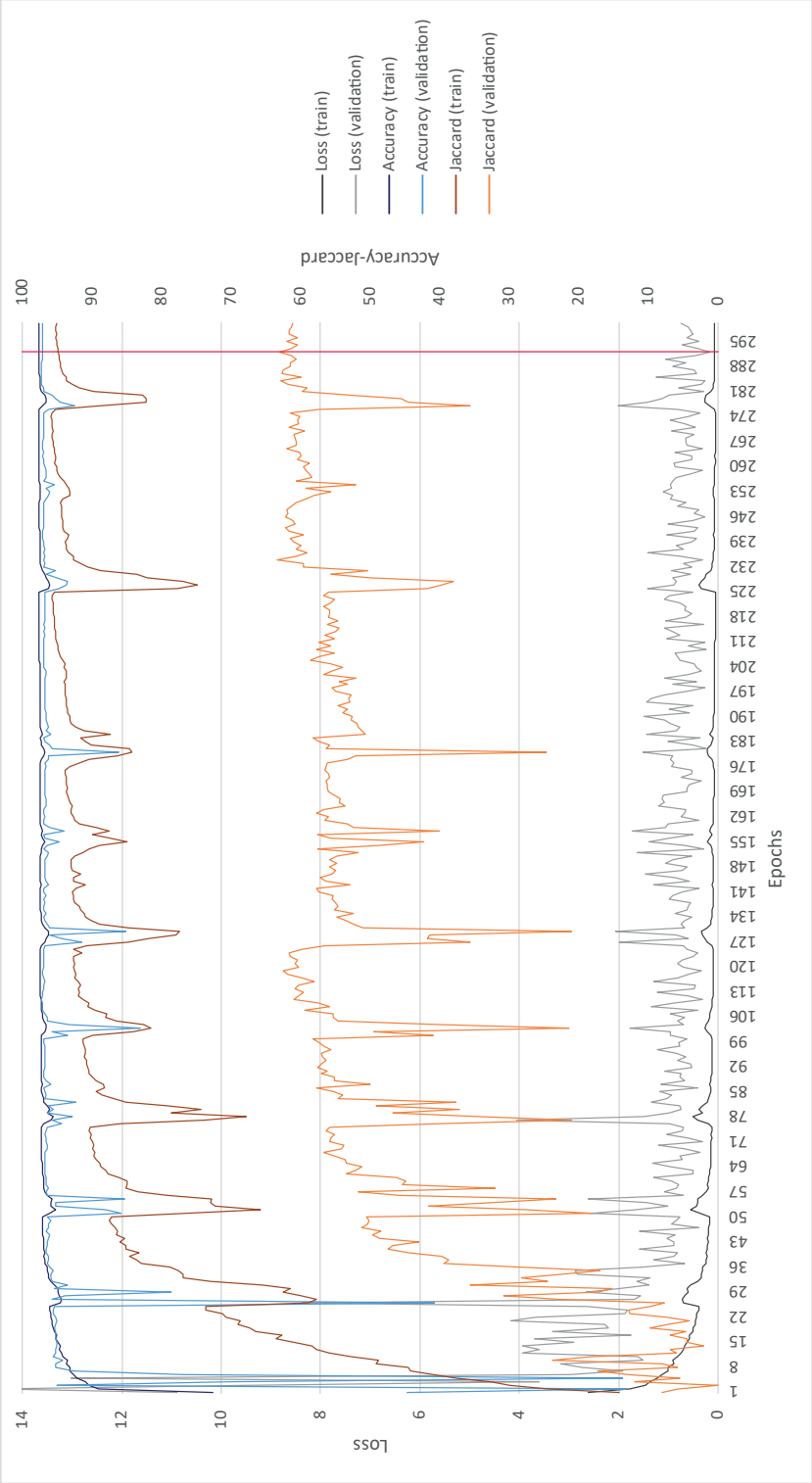
A.3.12 Pérdidas de entrenamiento y validación de LinkNet-VGG-16 para todas las funciones de pérdida



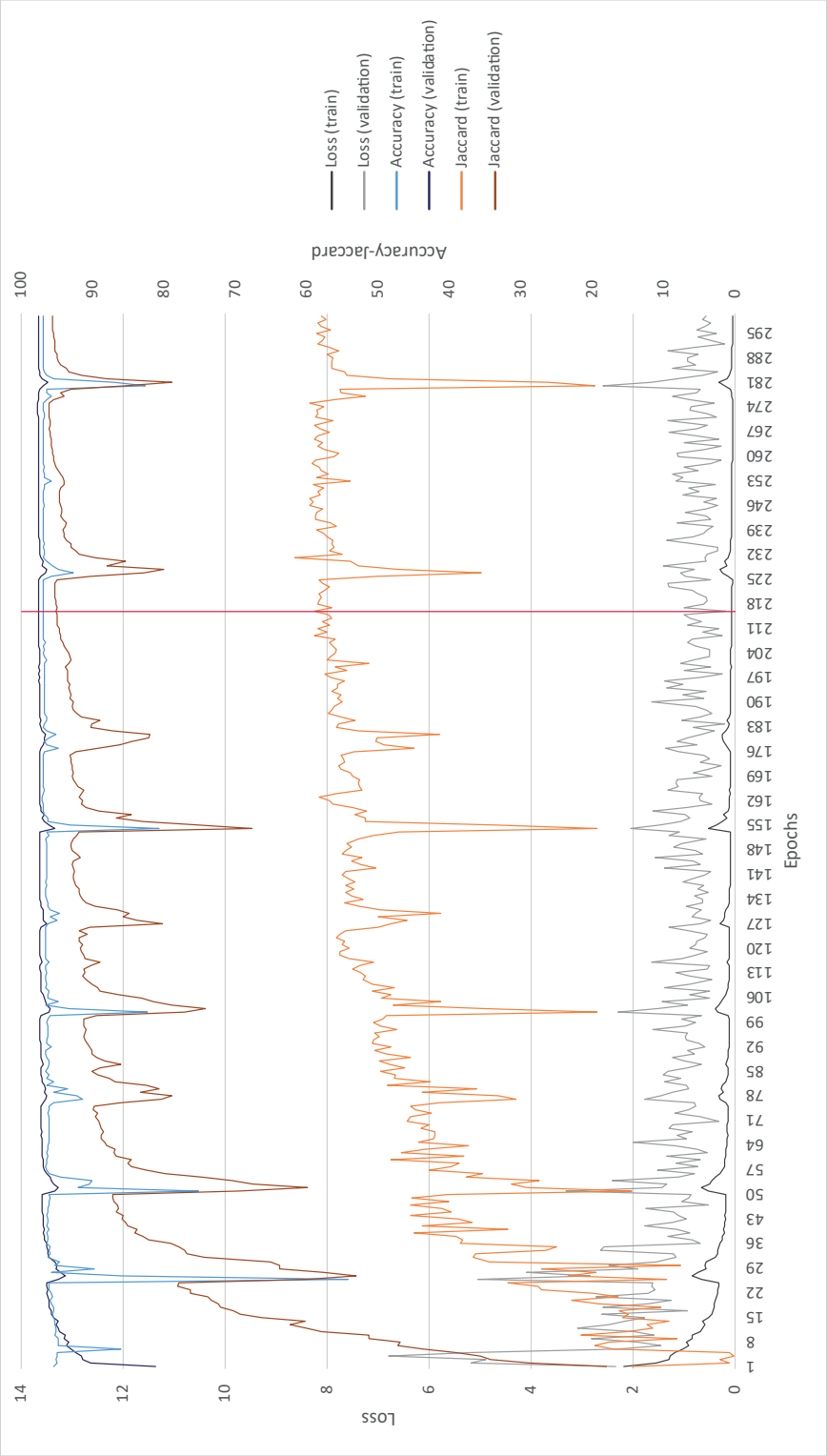
A.3.13 Pérdidas de entrenamiento y validación de LinkNet-Densenet121 para todas las funciones de pérdida



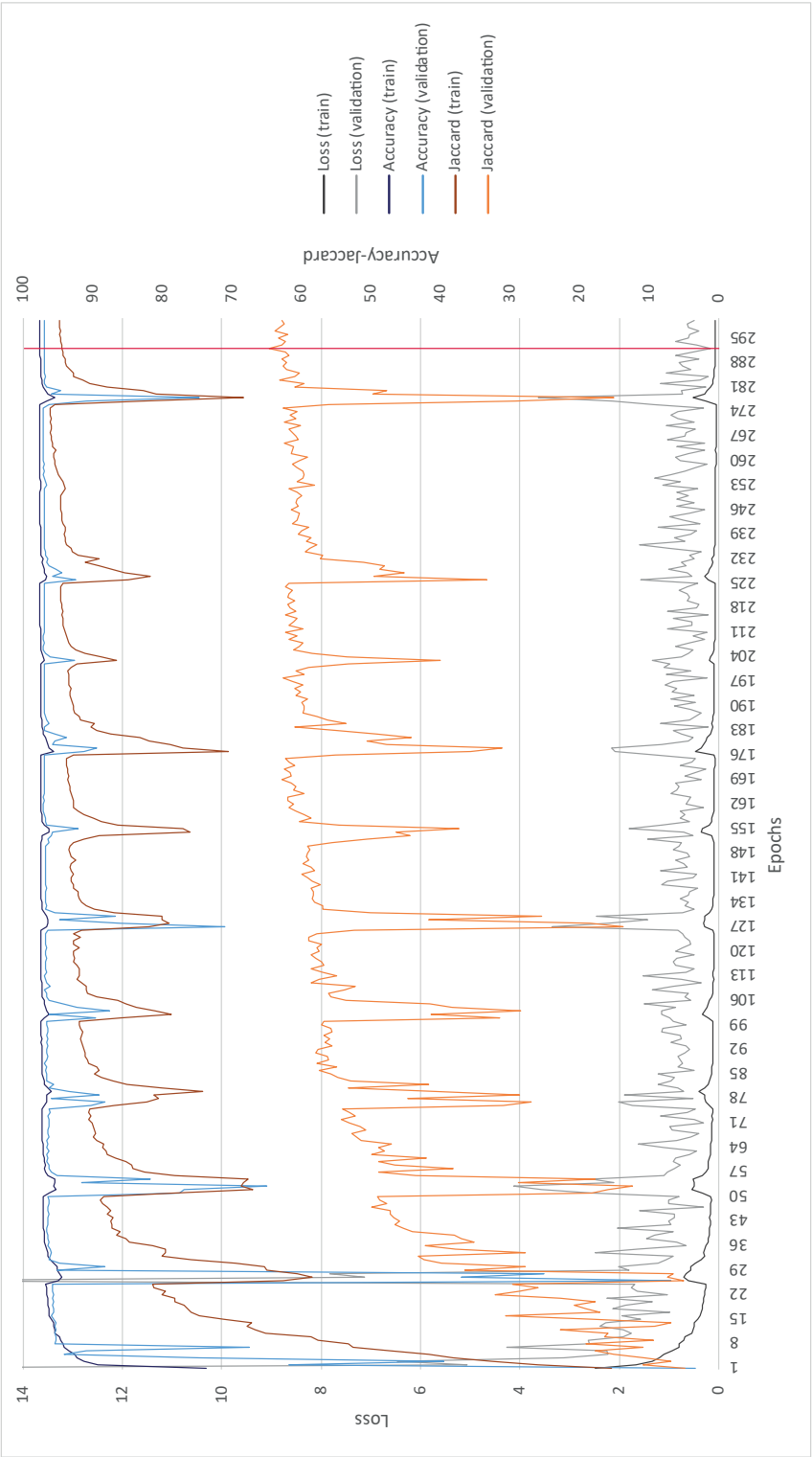
A.3.14 Pérdidas, exactitud e índice de Jaccard durante el entrenamiento para los conjuntos de entrenamiento y validación de U-Net-VGG-16



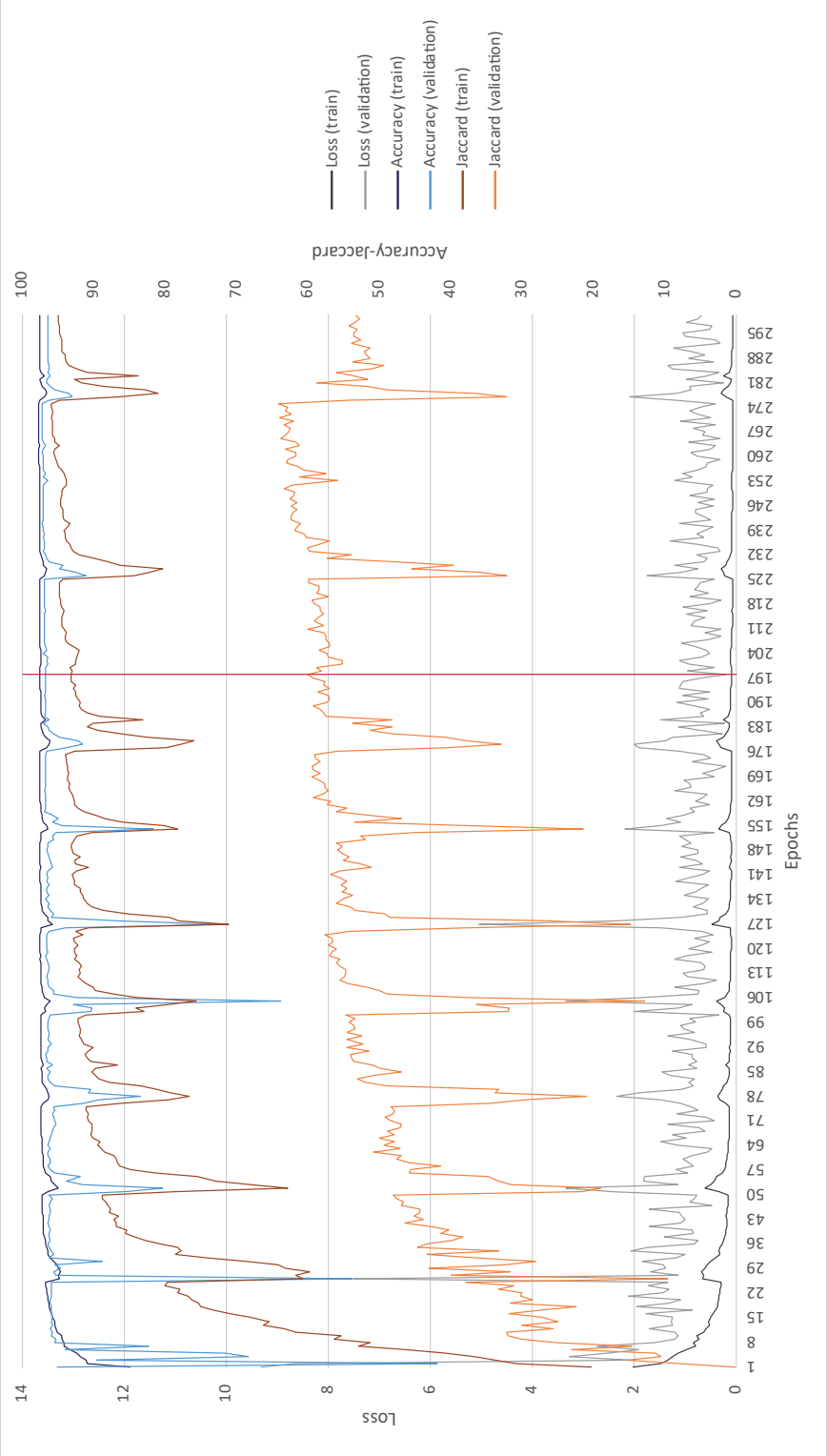
Loss rate, accuracy and Jaccard during training for the train and validation sets using the Jaccard loss function.
The red line indicates the best epoch.



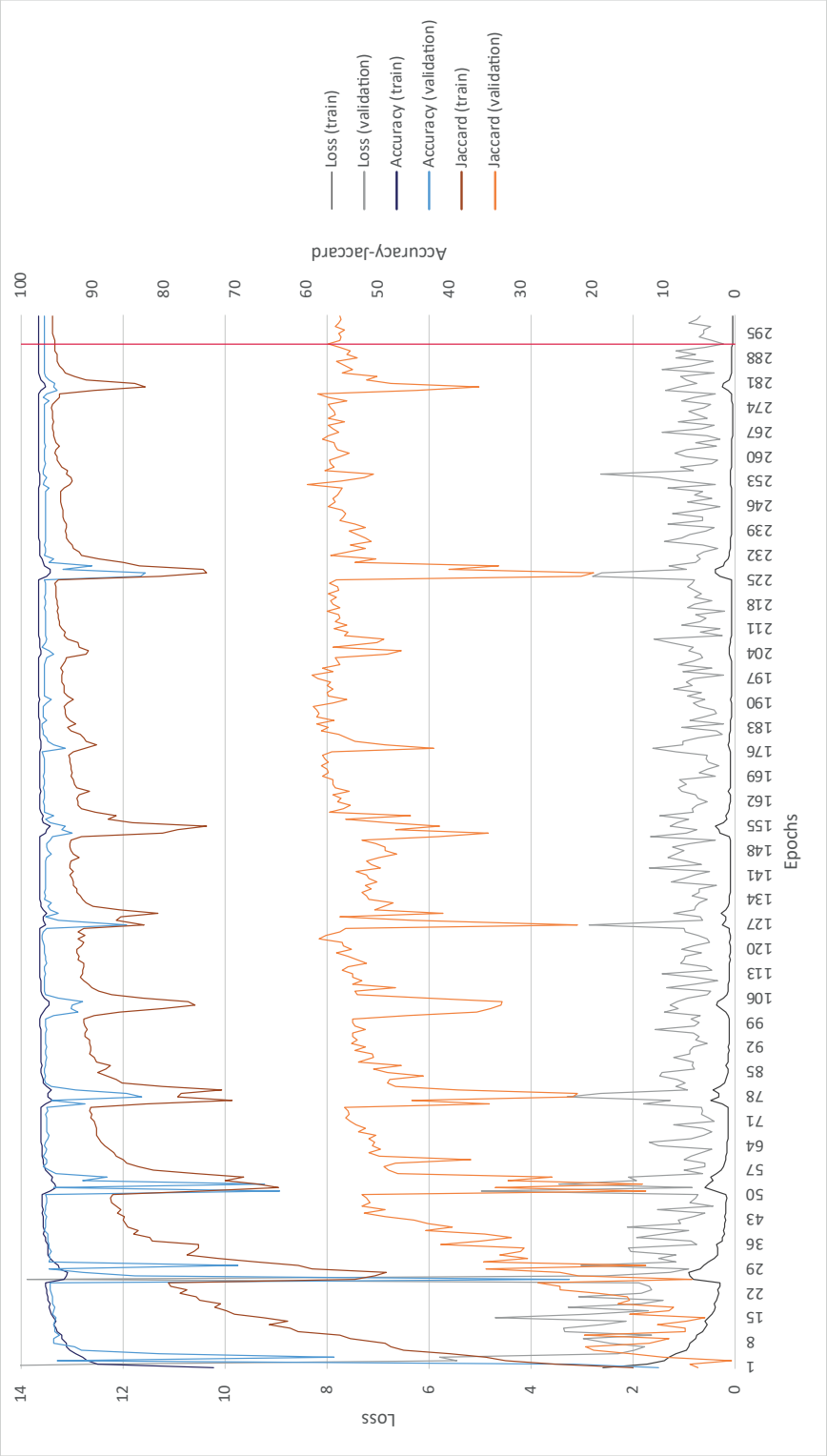
Loss rate, accuracy and Jaccard during training for the train and validation sets using the Dice loss function.
The red line indicates the best epoch.



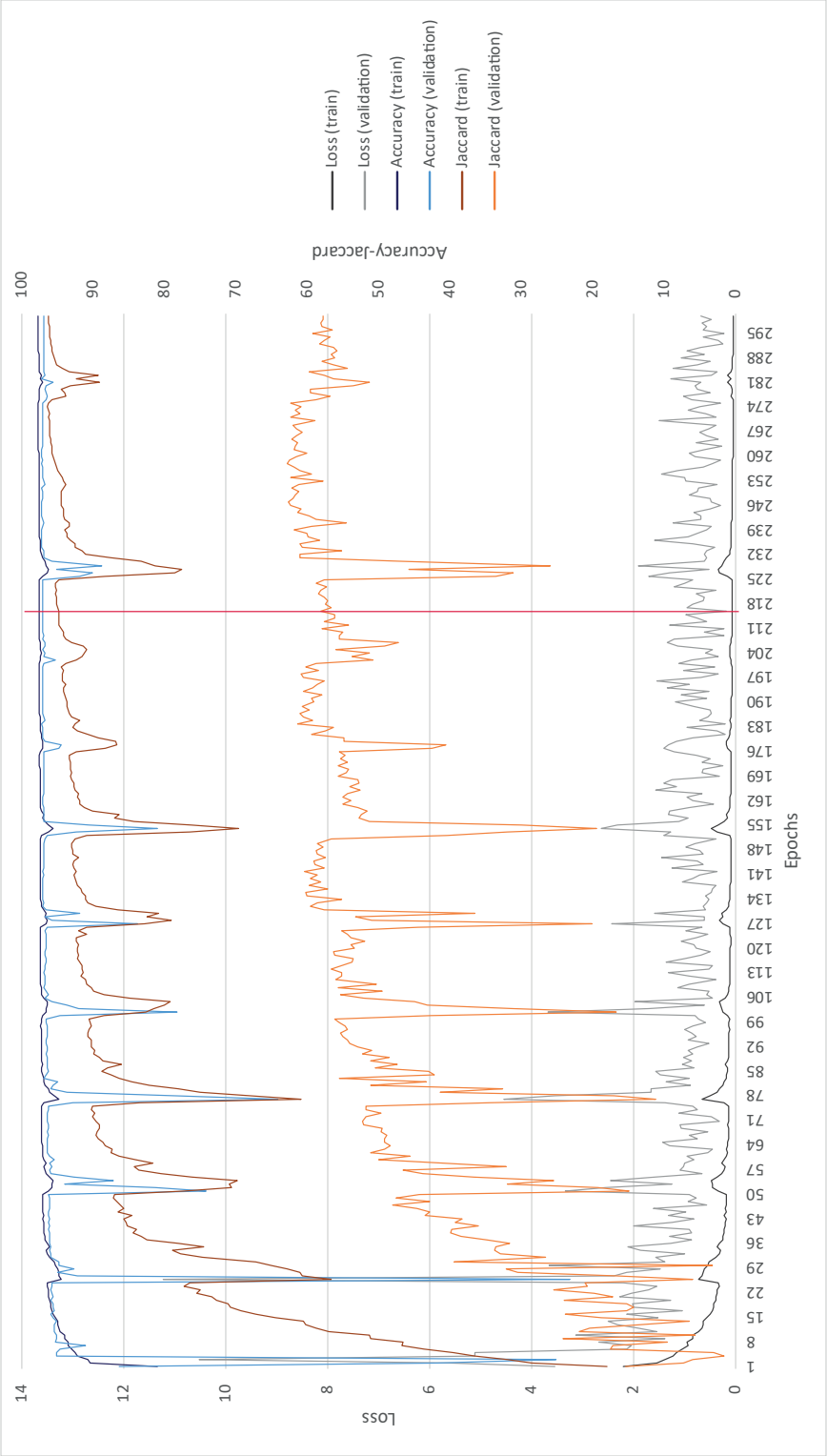
Loss rate, accuracy and Jaccard during training for the train and validation sets using the binary cross entropy loss function. The red line indicates the best epoch.



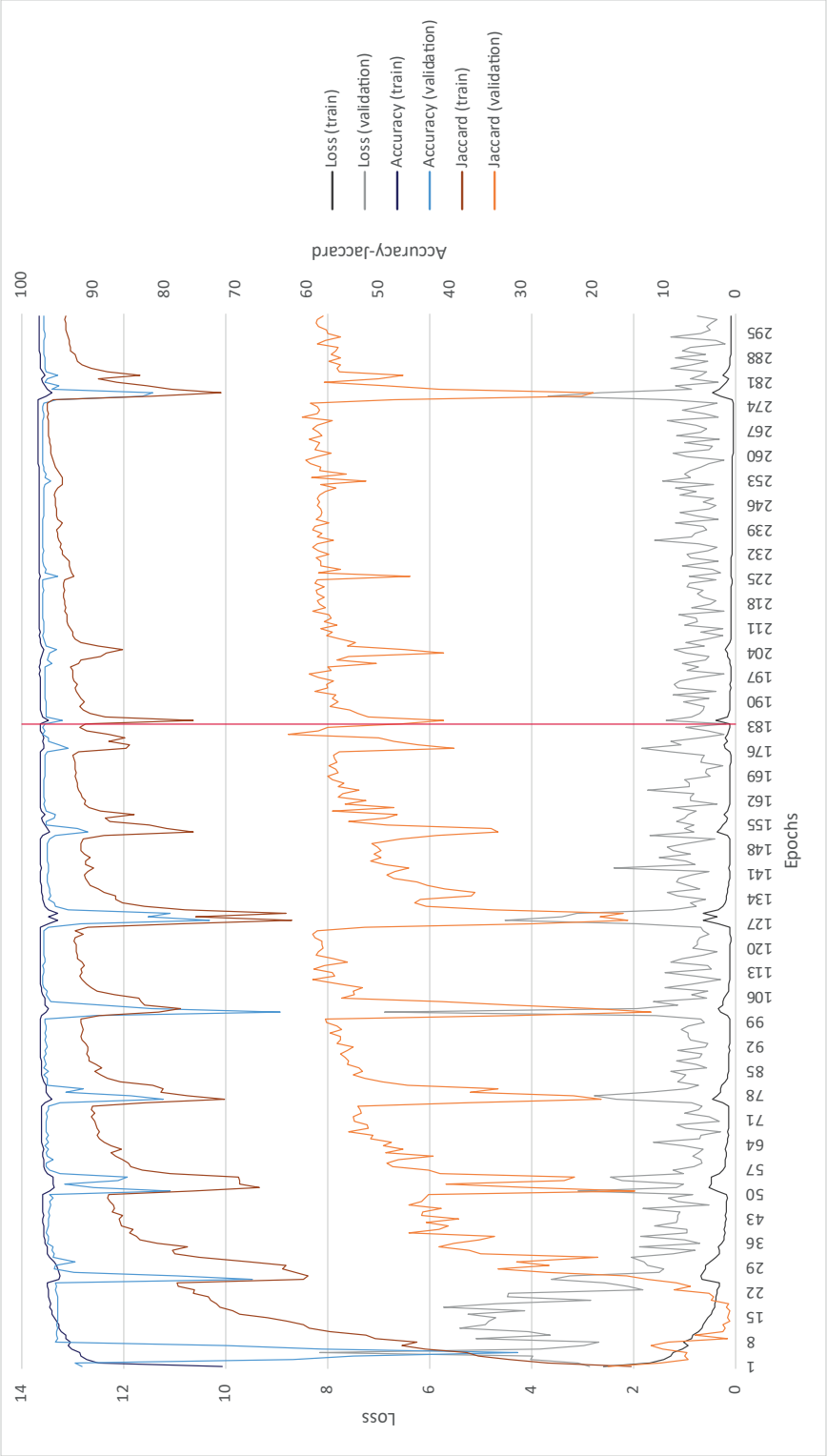
Loss rate, accuracy and Jaccard during training for the train and validation sets using the binary focal loss function. The red line indicates the best epoch.



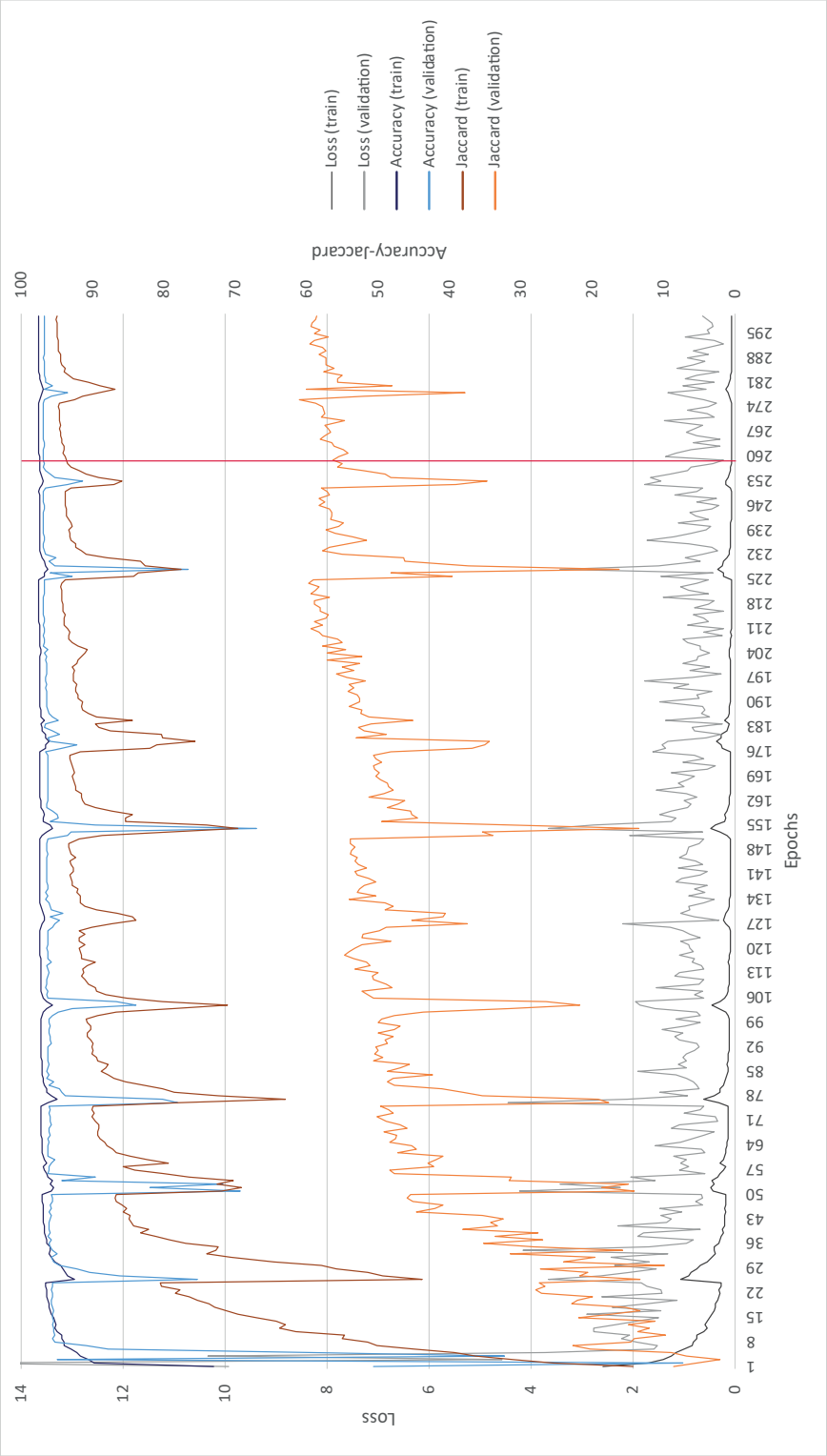
Loss rate, accuracy and Jaccard during training for the train and validation sets using the Tversky loss function.
The red line indicates the best epoch.



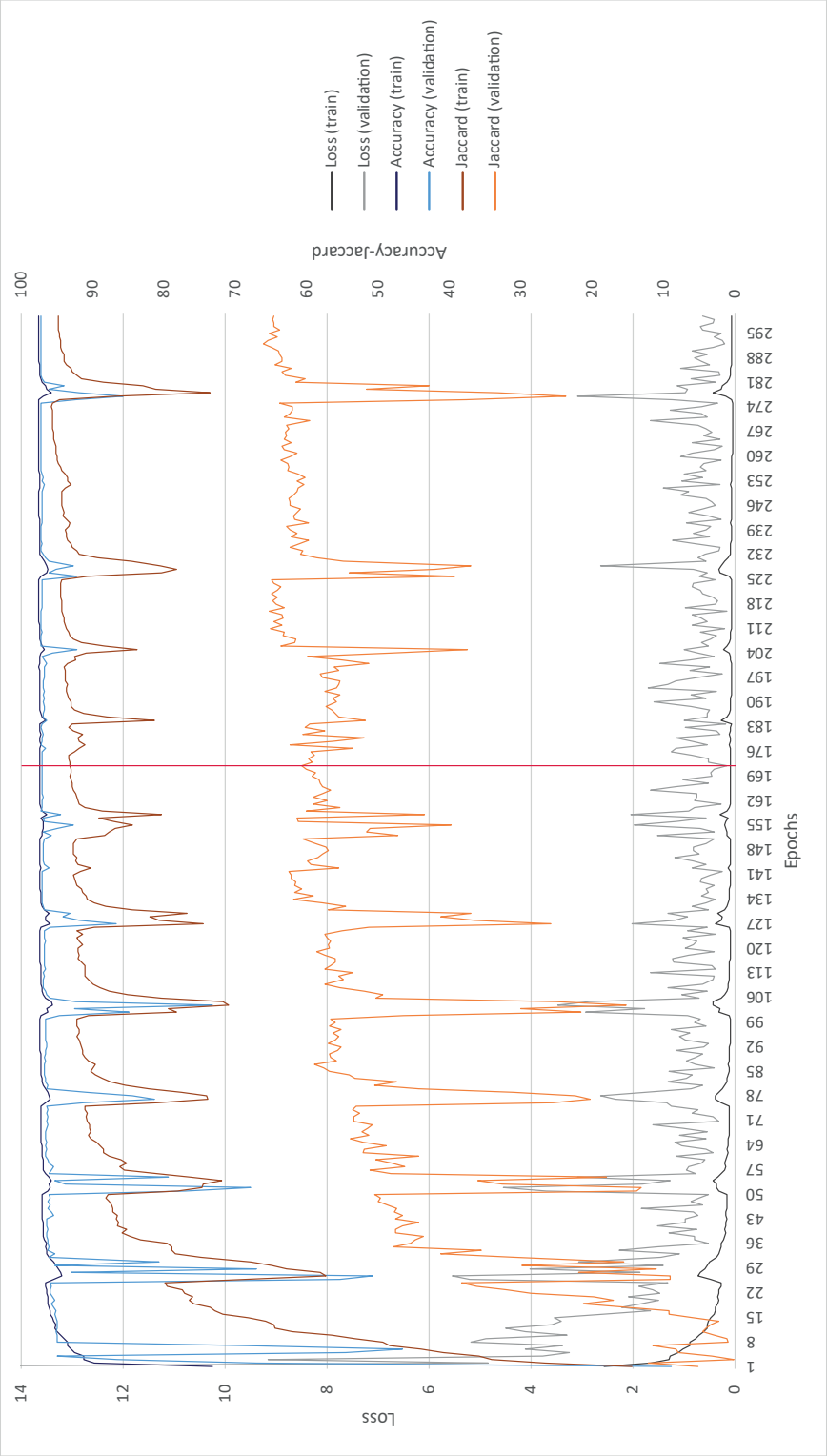
Loss rate, accuracy and Jaccard during training for the train and validation sets using the focal Tversky loss function. The red line indicates the best epoch.



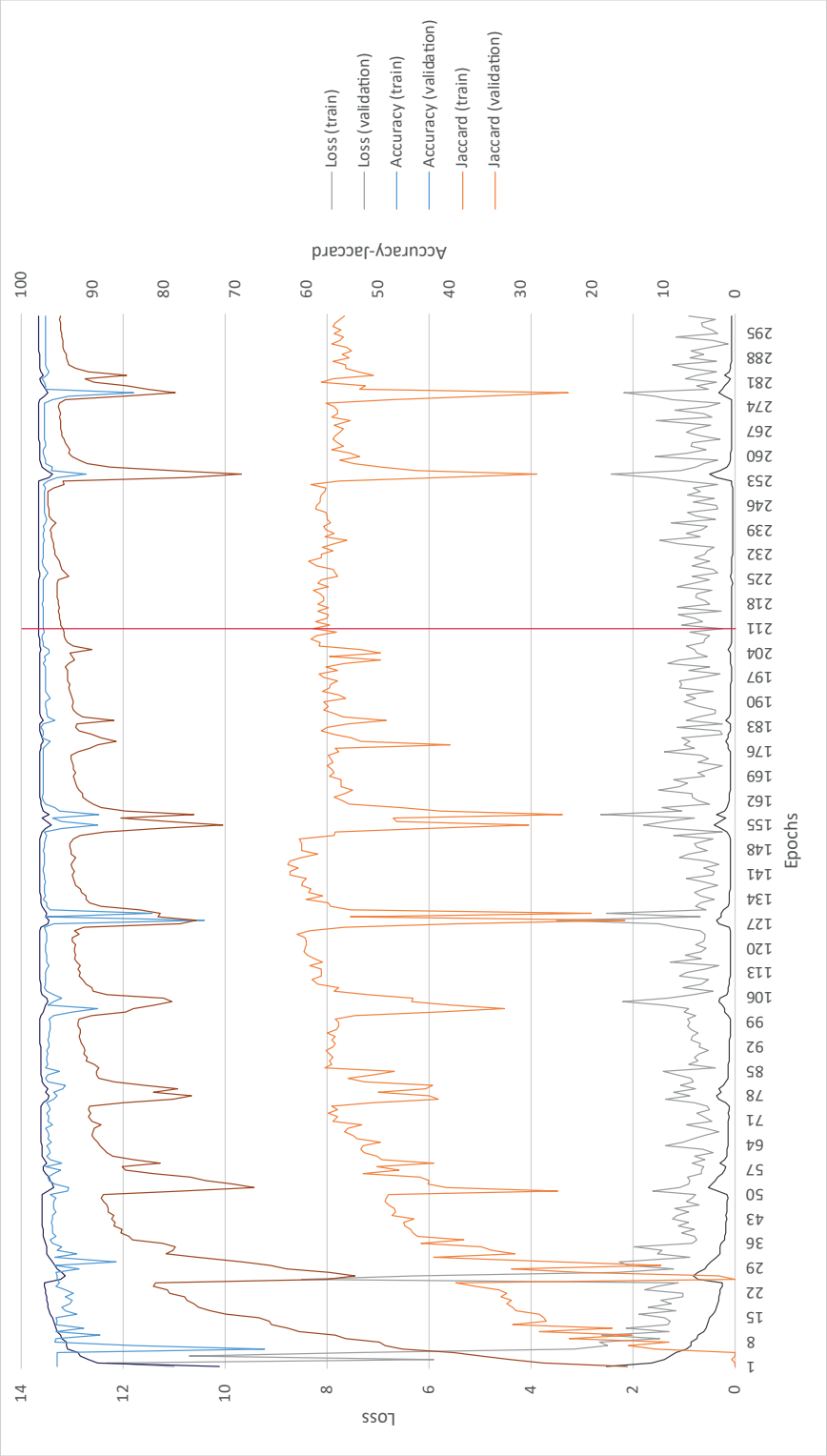
Loss rate, accuracy and Jaccard during training for the train and validation sets using the Lovász-Hinge loss function. The red line indicates the best epoch.



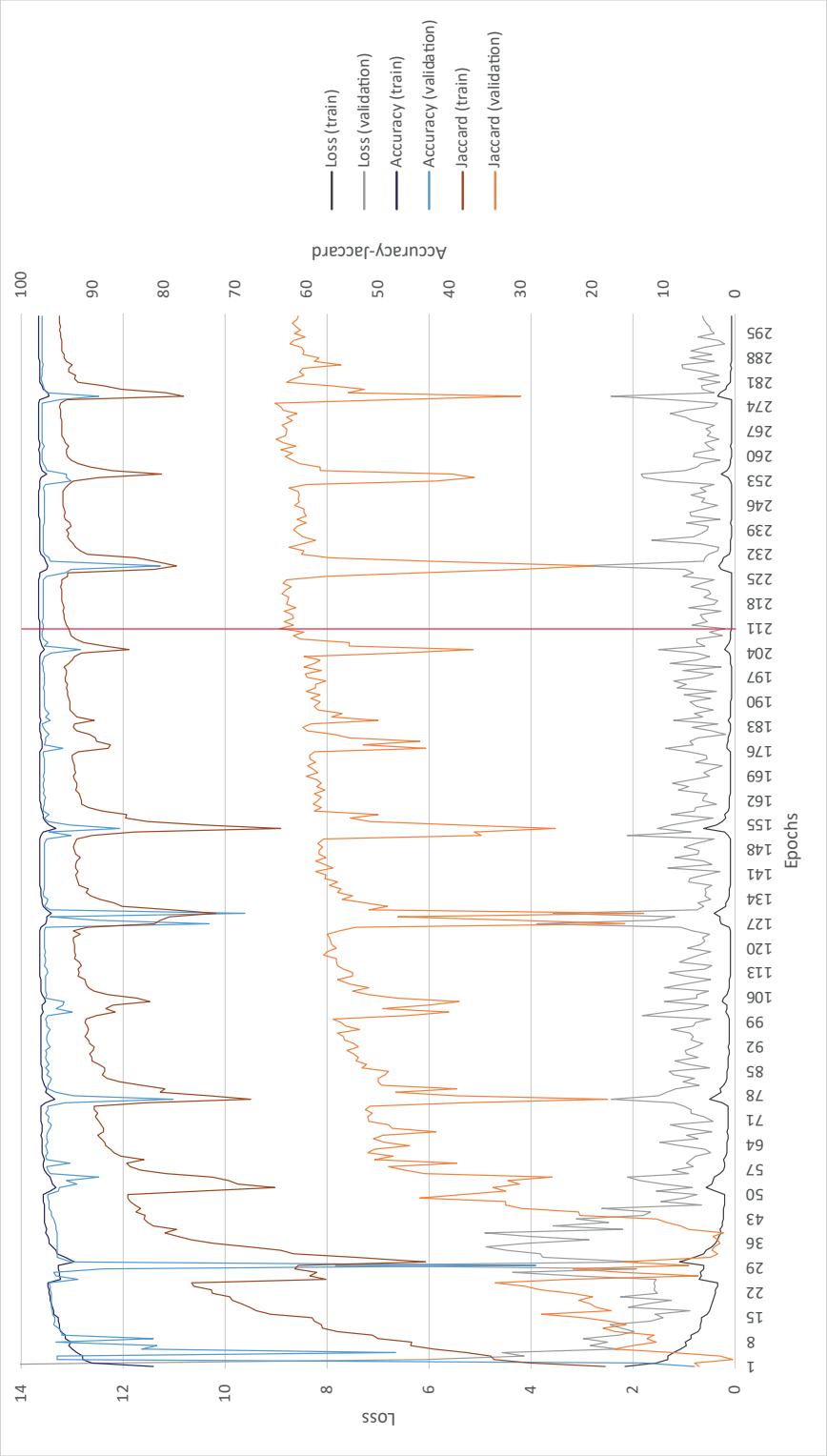
Loss rate, accuracy and Jaccard during training for the train and validation sets using the sum loss function. The red line indicates the best epoch.



Loss rate, accuracy and Jaccard during training for the train and validation sets using the mean loss function.
The red line indicates the best epoch.

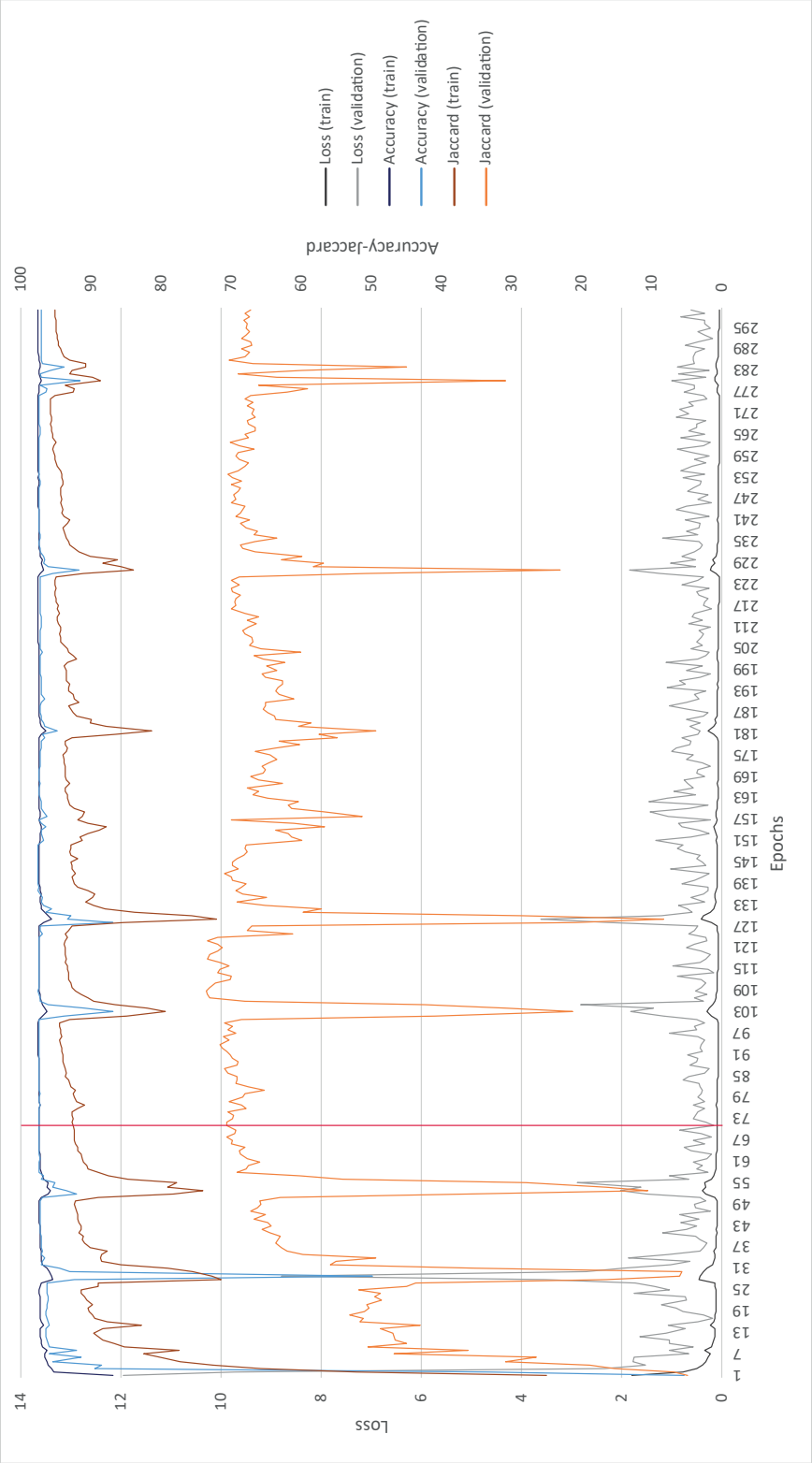


Loss rate, accuracy and Jaccard during training for the train and validation sets using the eigenloss function.
The red line indicates the best epoch.

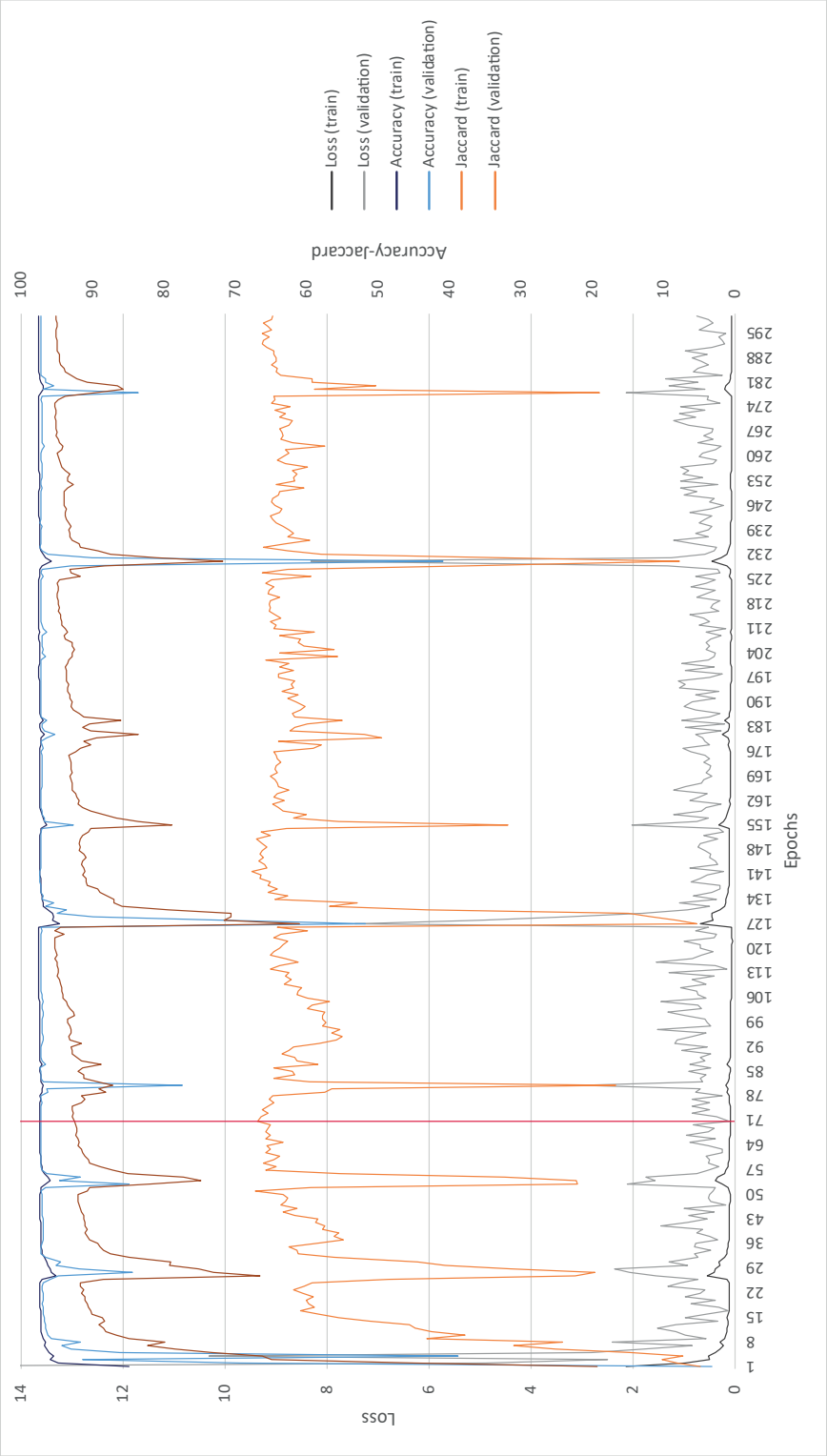


Loss rate, accuracy and Jaccard during training for the train and validation sets using the normalized eigenloss function. The red line indicates the best epoch.

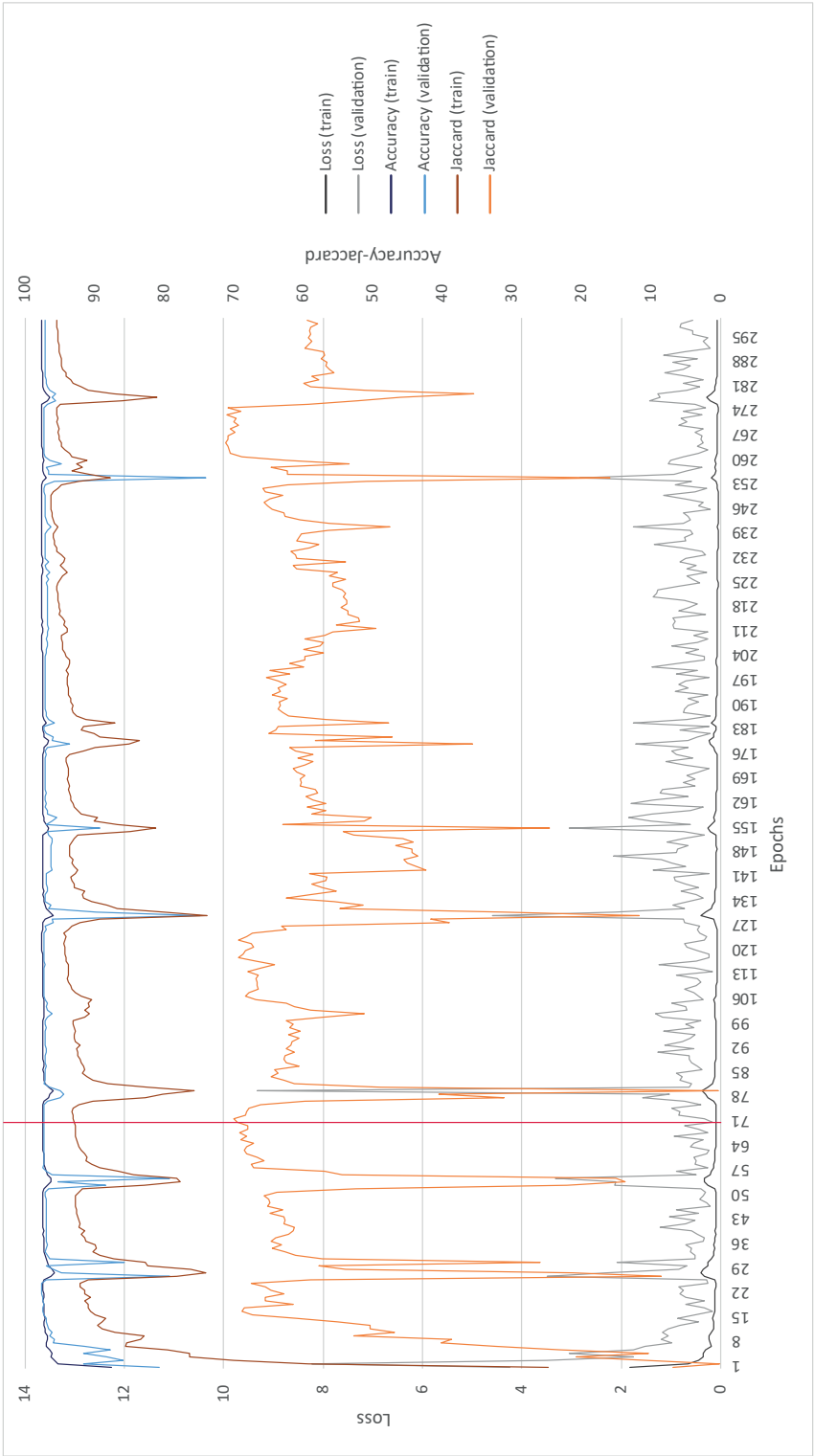
A.3.15 Pérdidas, exactitud e índice de Jaccard durante el entrenamiento para los conjuntos de entrenamiento y validación de U-Net-Densenet121



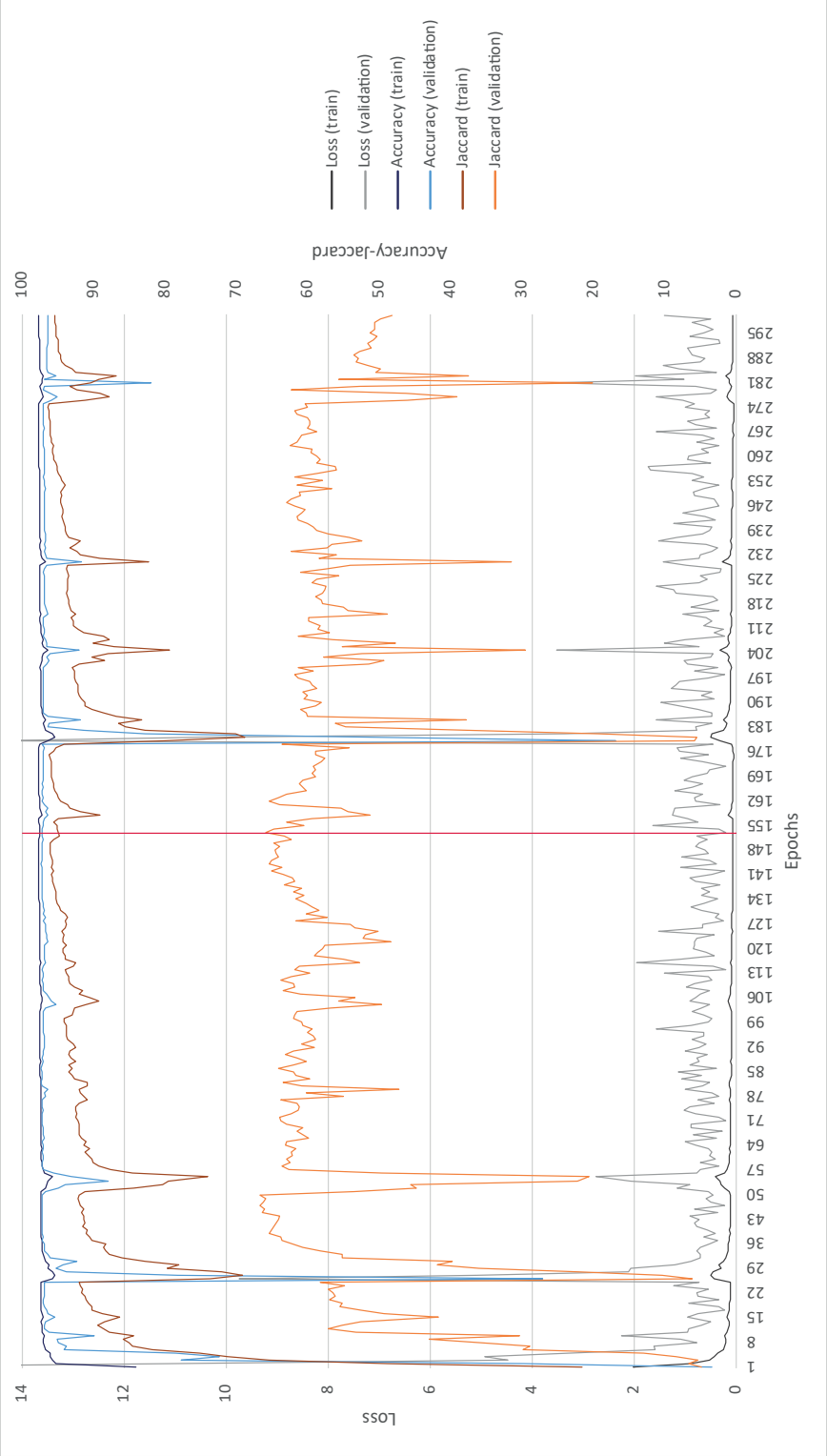
Loss rate, accuracy and Jaccard during training for the train and validation sets using the Jaccard loss function.
The red line indicates the best epoch.



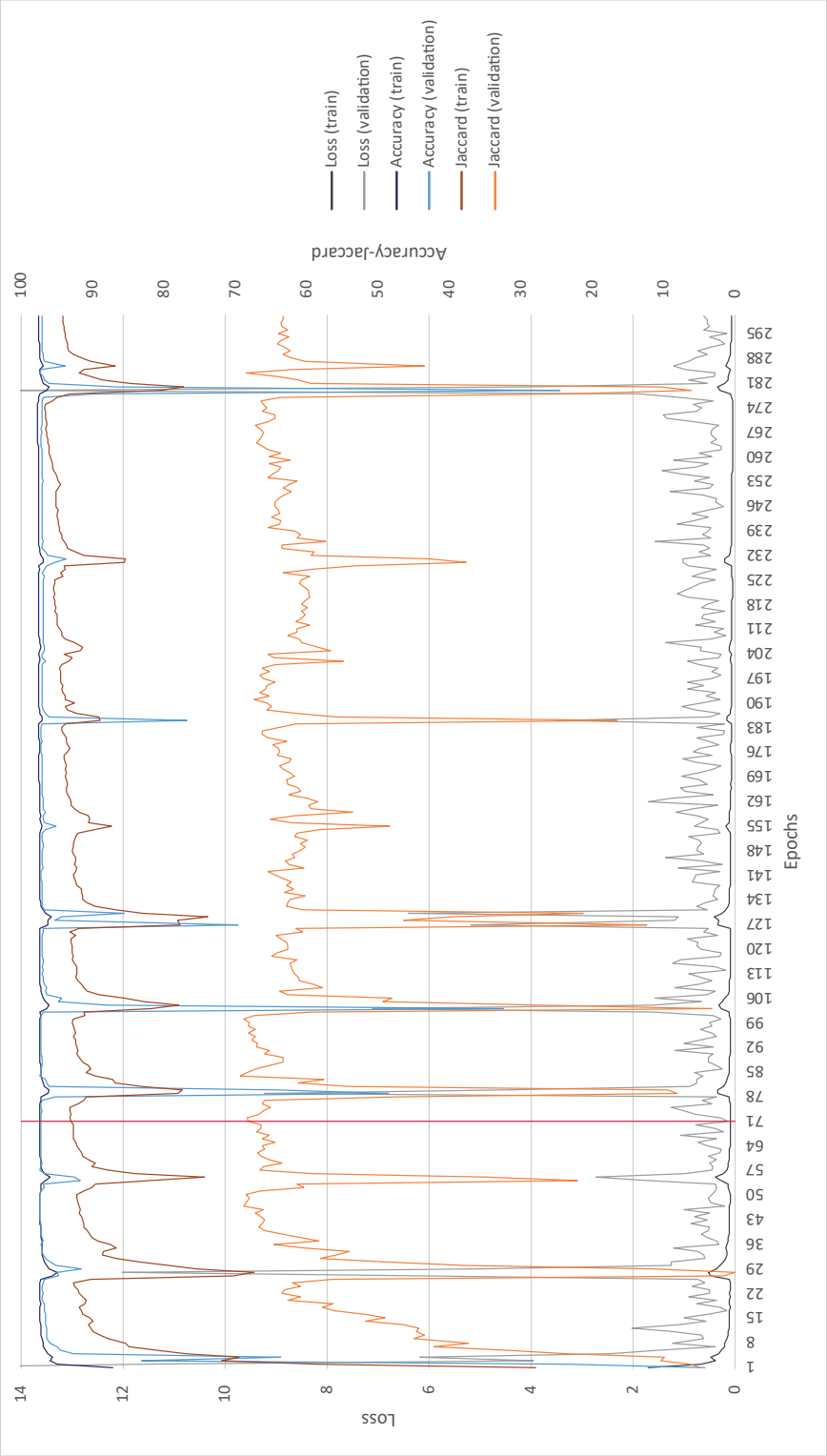
Loss rate, accuracy and Jaccard during training for the train and validation sets using the Dice loss function.
The red line indicates the best epoch.



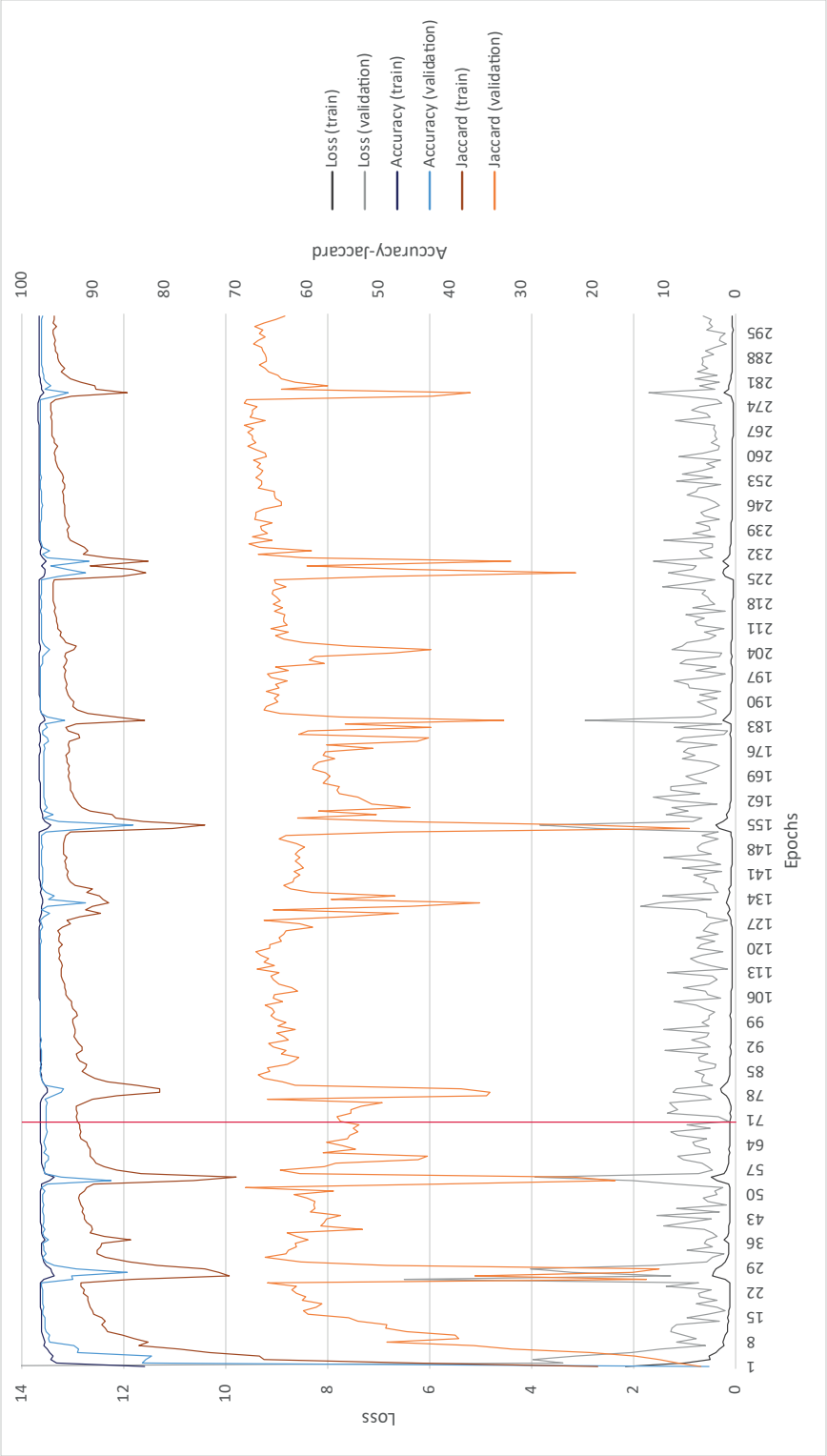
Loss rate, accuracy and Jaccard during training for the train and validation sets using the binary cross entropy loss function. The red line indicates the best epoch.



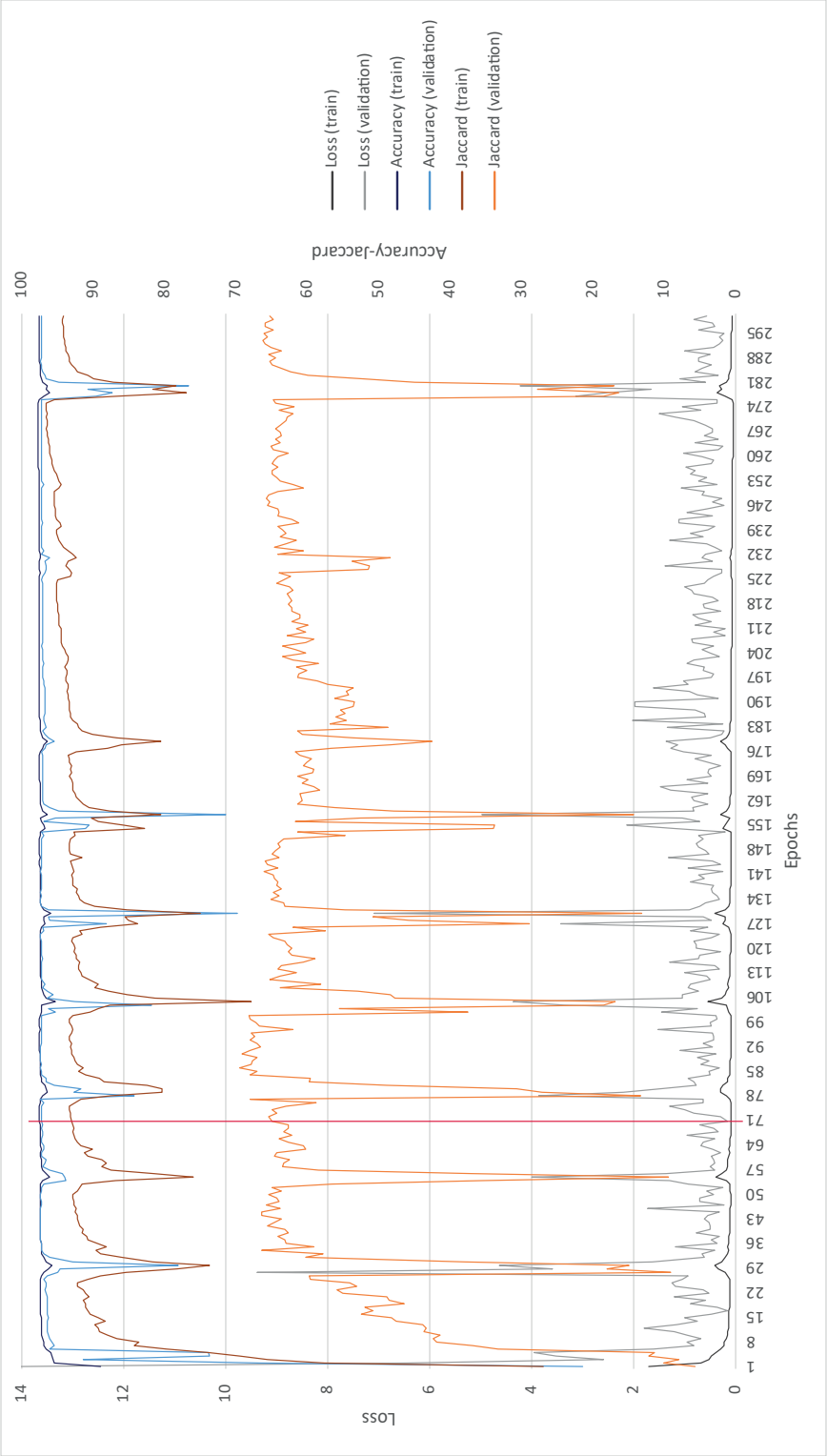
Loss rate, accuracy and Jaccard during training for the train and validation sets using the binary focal loss function. The red line indicates the best epoch.



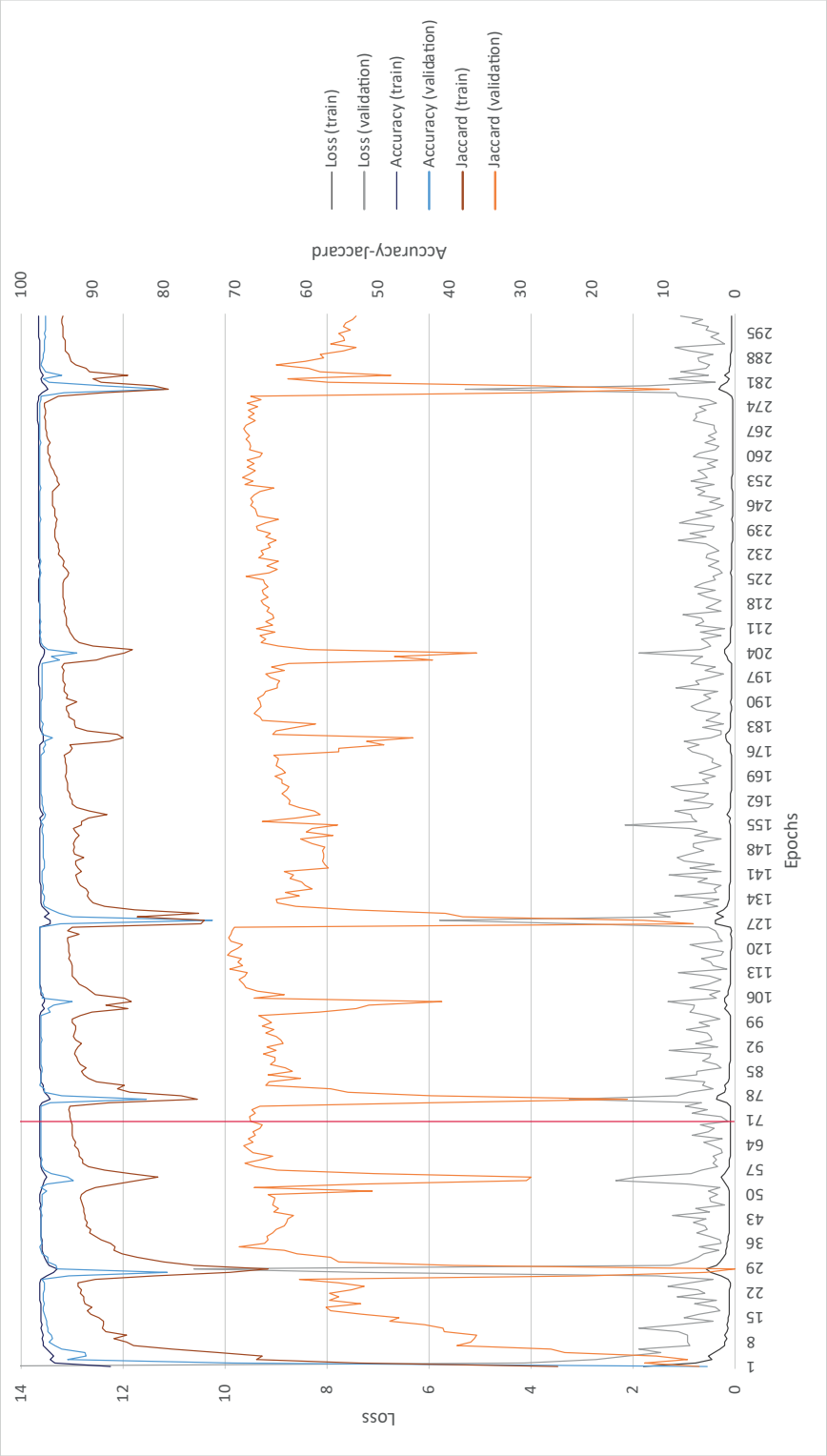
Loss rate, accuracy and Jaccard during training for the train and validation sets using the Tversky loss function



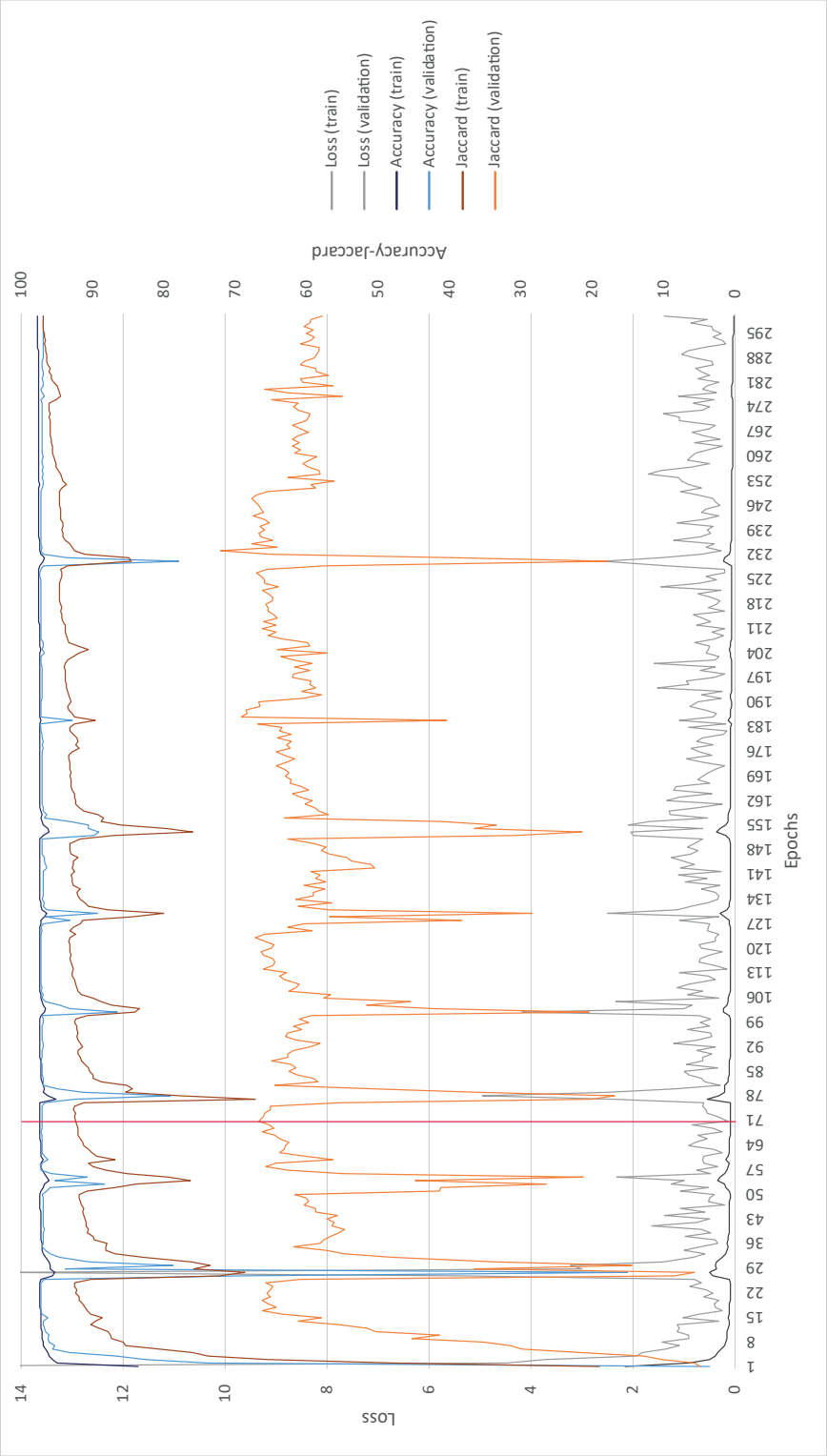
Loss rate, accuracy and Jaccard during training for the train and validation sets using the focal Tversky loss function. The red line indicates the best epoch.



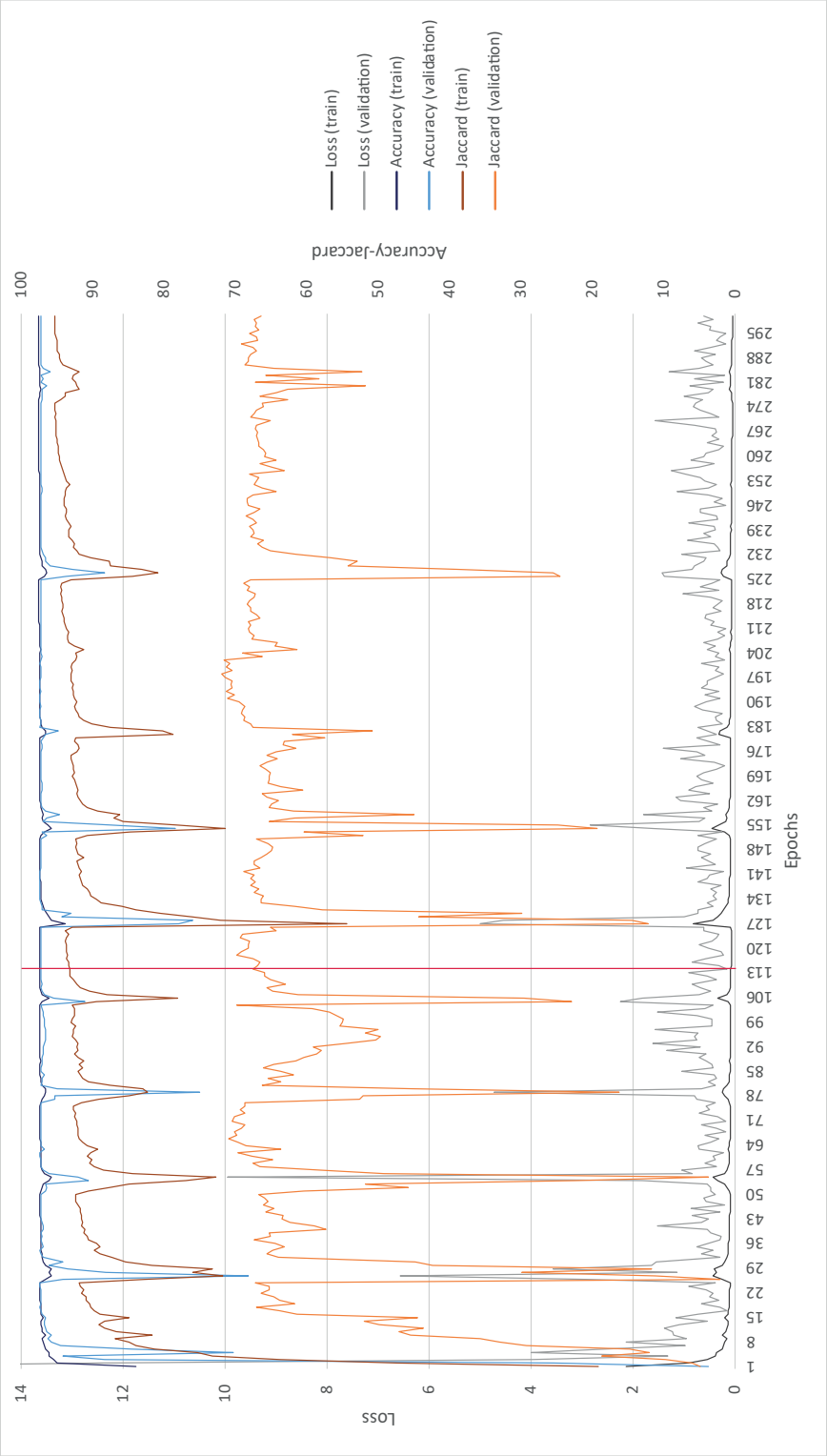
Loss rate, accuracy and Jaccard during training for the train and validation sets using the Lovász-Hinge loss function. The red line indicates the best epoch.



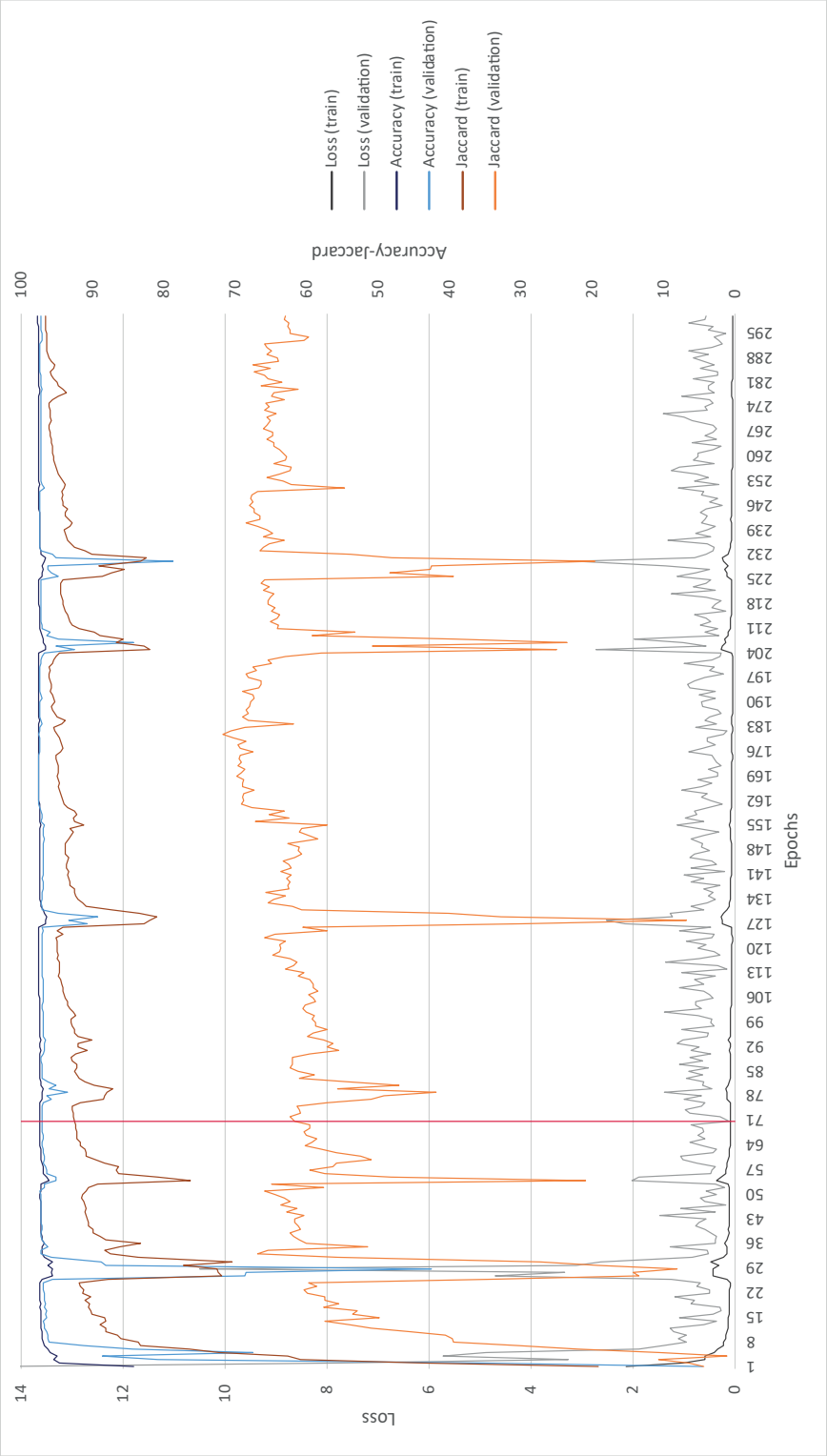
Loss rate, accuracy and Jaccard during training for the train and validation sets using the sum loss function. The red line indicates the best epoch.



Loss rate, accuracy and Jaccard during training for the train and validation sets using the mean loss function

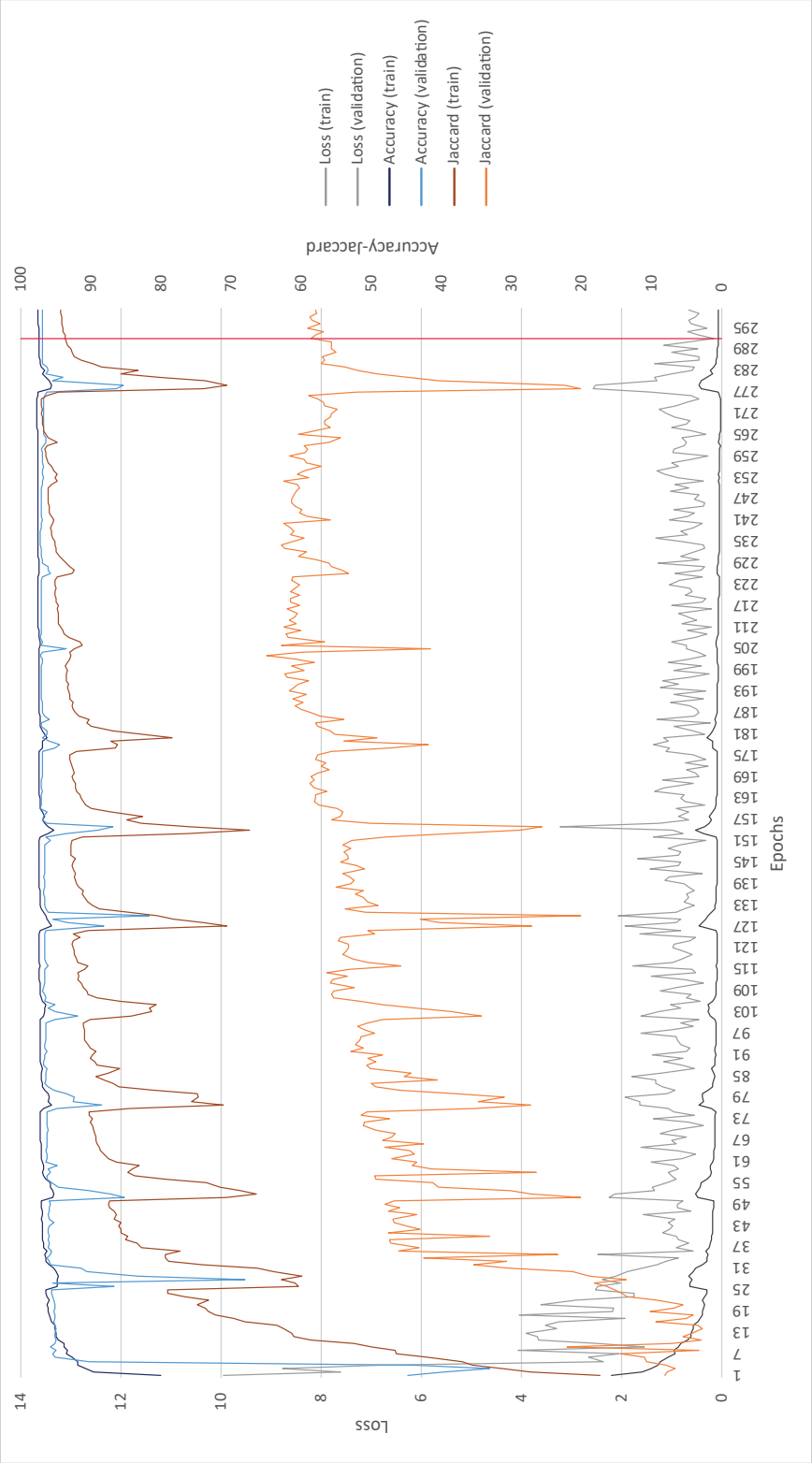


Loss rate, accuracy and Jaccard during training for the train and validation sets using the eigenloss function.
The red line indicates the best epoch.

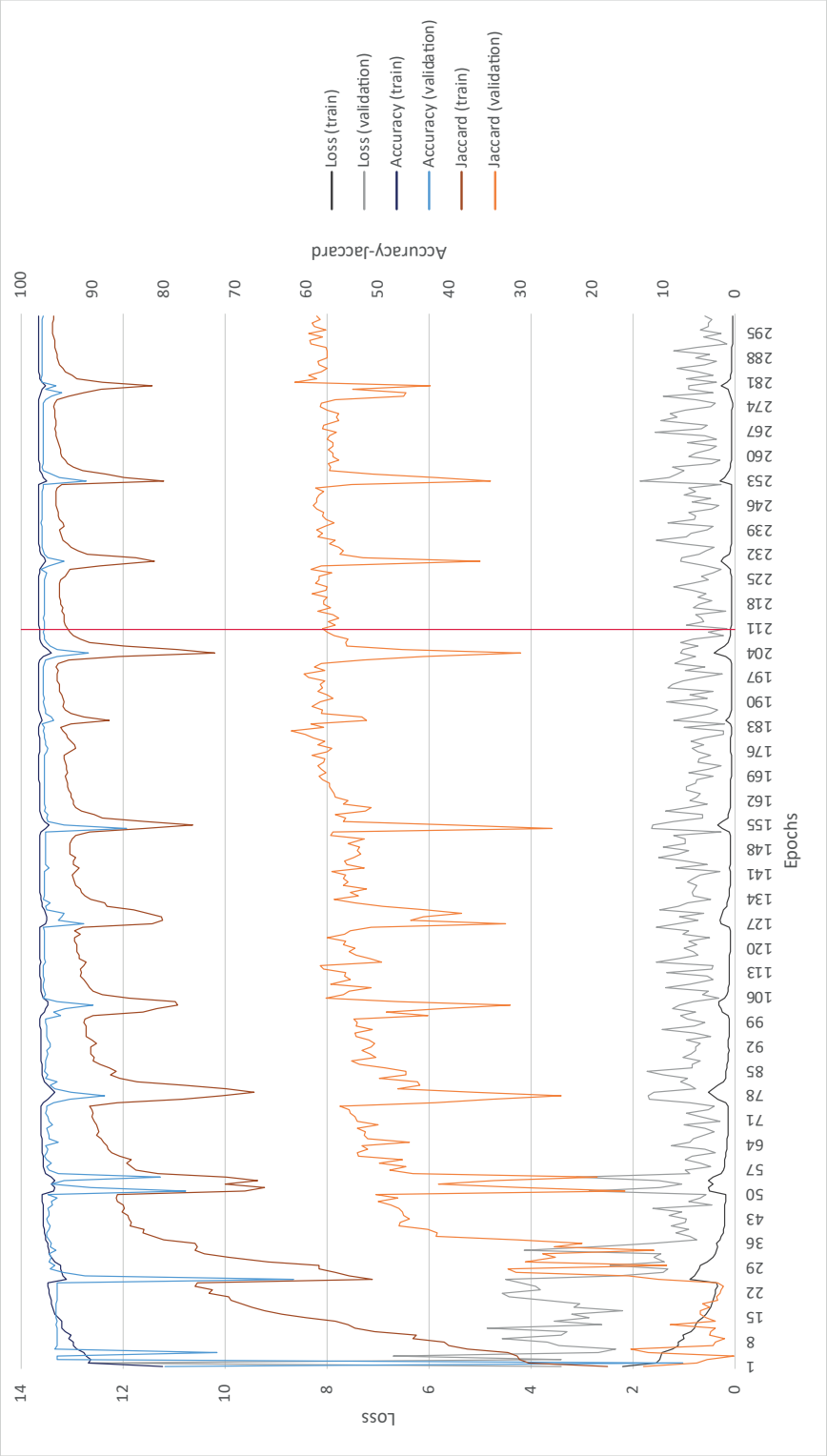


Loss rate, accuracy and Jaccard during training for the train and validation sets using the normalized eigenloss function. The red line indicates the best epoch.

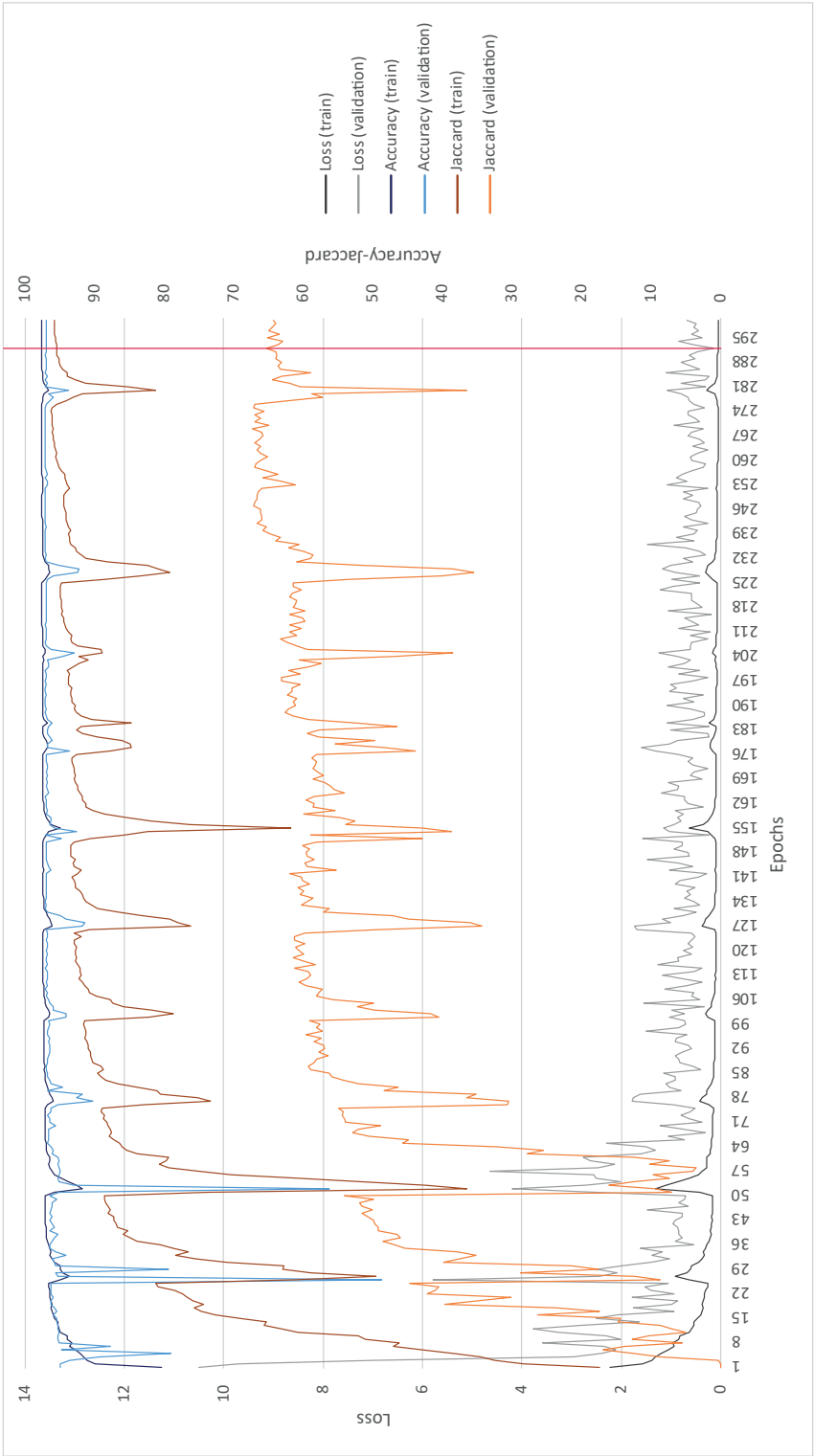
A.3.16 Pérdidas, exactitud e índice de Jaccard durante el entrenamiento para los conjuntos de entrenamiento y validación de LinkNet-VGG-16



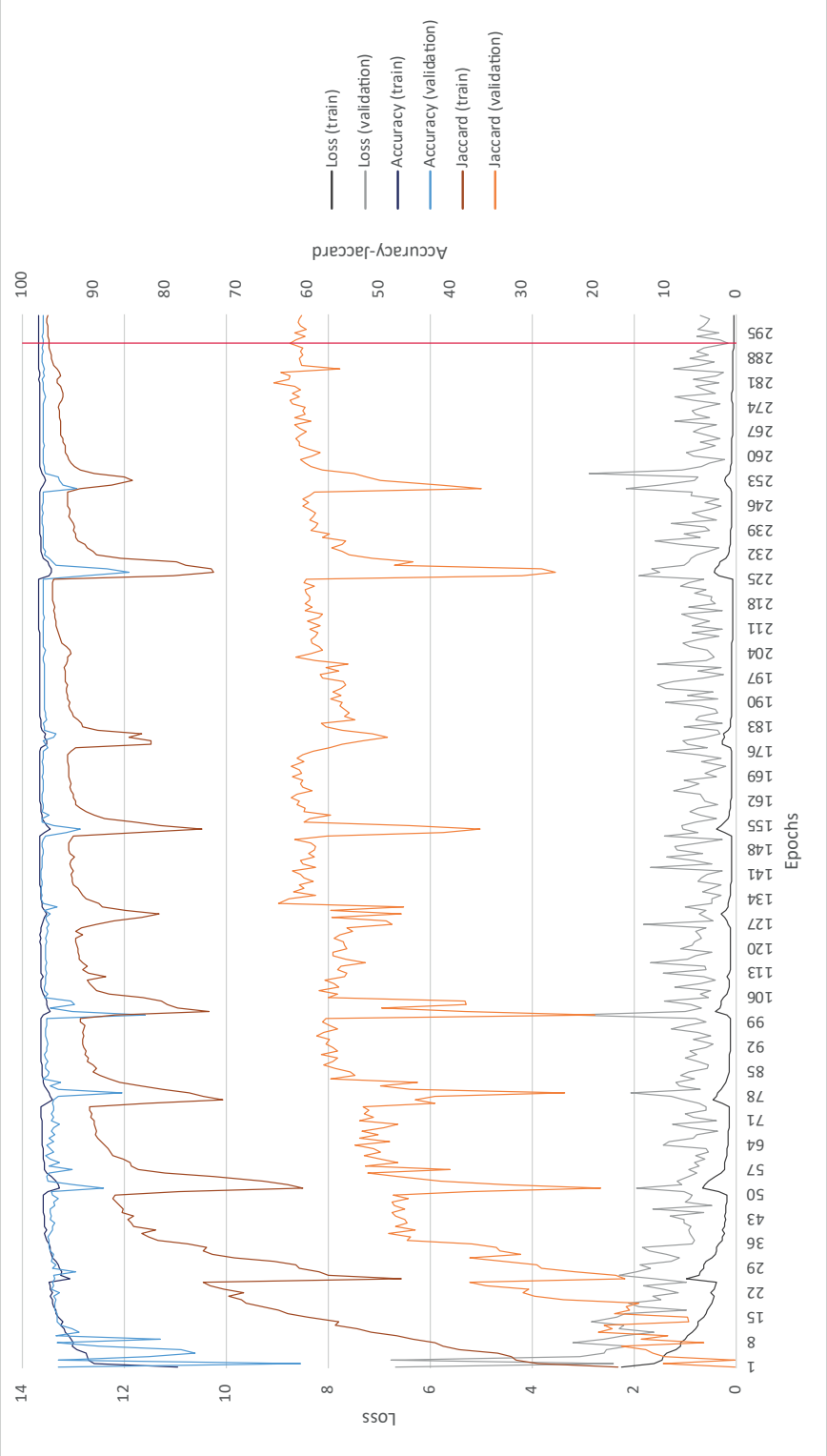
Loss rate, accuracy and Jaccard during training for the train and validation sets using the Jaccard loss function.
The red line indicates the best epoch.



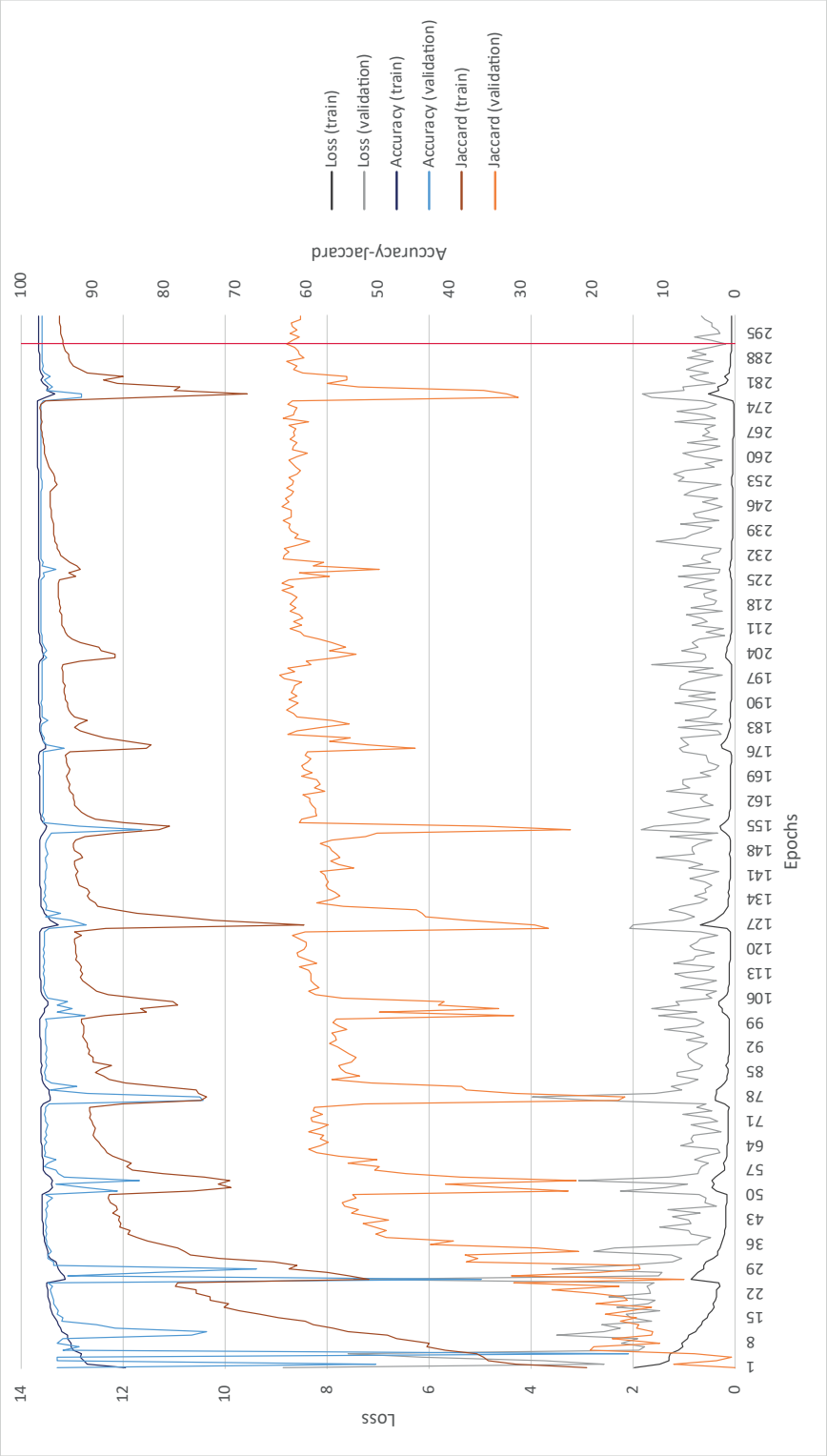
Loss rate, accuracy and Jaccard during training for the train and validation sets using the Dice loss function.
The red line indicates the best epoch.



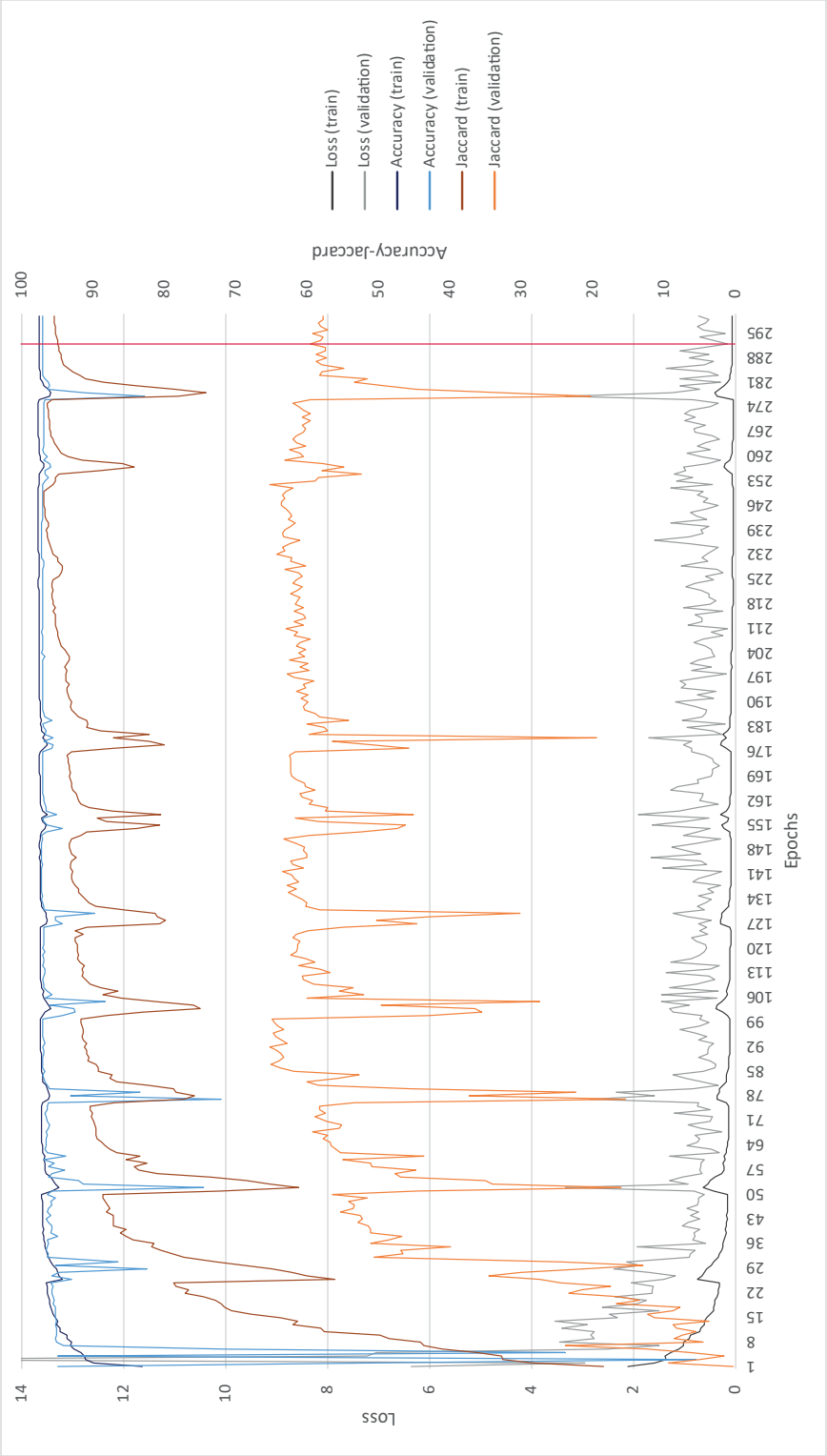
Loss rate, accuracy and Jaccard during training for the train and validation sets using the binary cross entropy loss function. The red line indicates the best epoch.



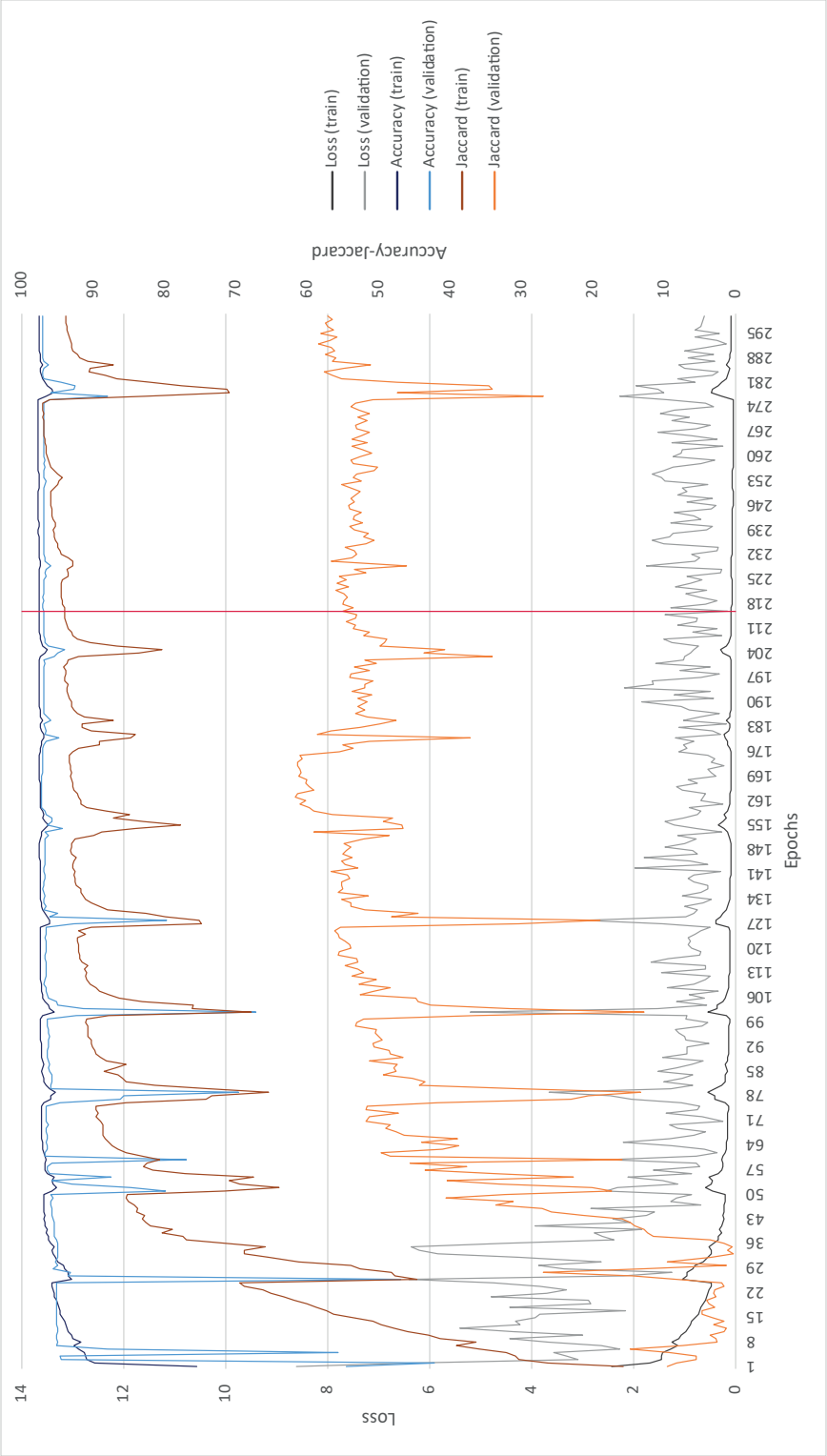
Loss rate, accuracy and Jaccard during training for the train and validation sets using the binary focal loss function



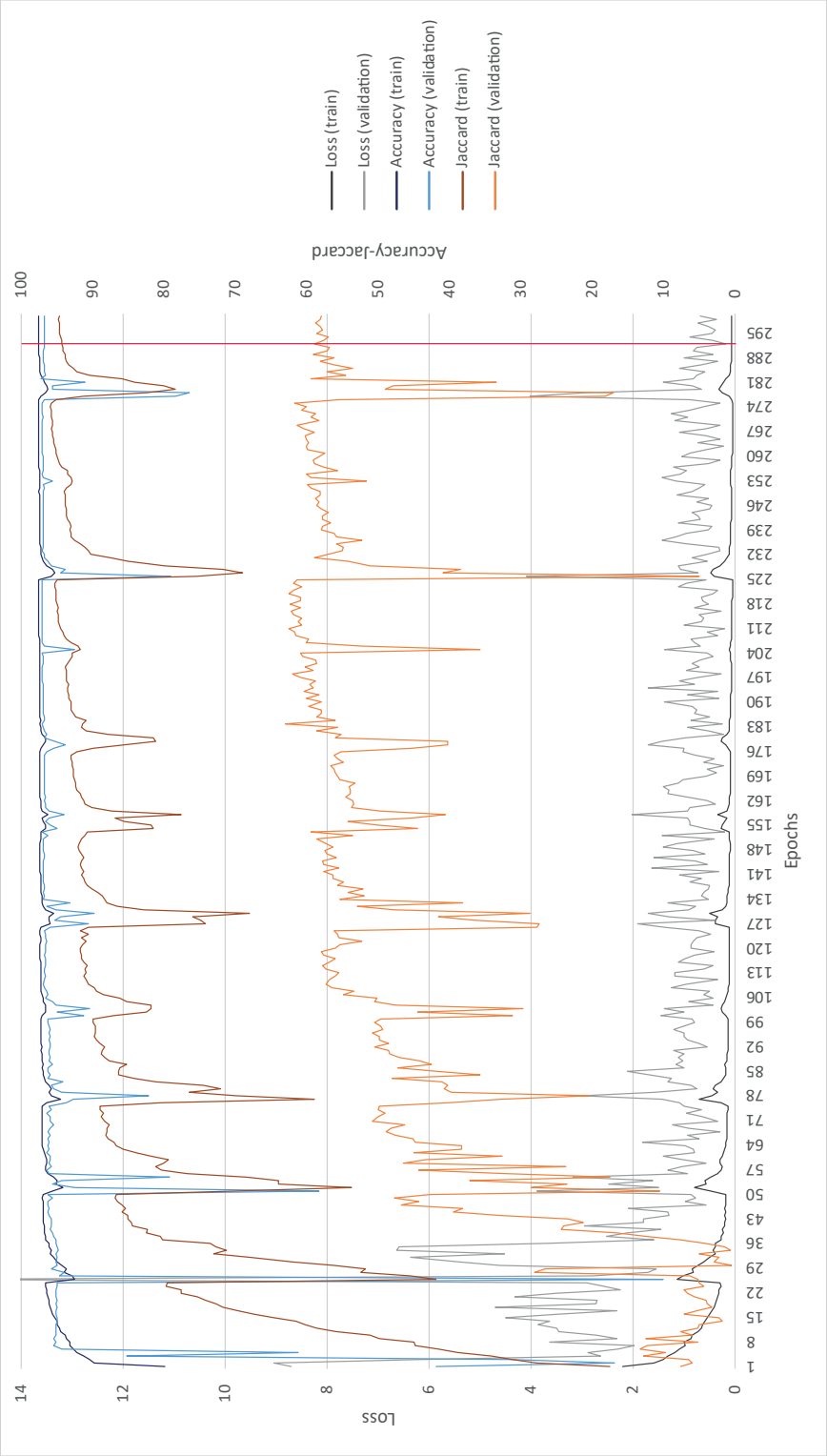
Loss rate, accuracy and Jaccard during training for the train and validation sets using the Tversky loss function. The red line indicates the best epoch.



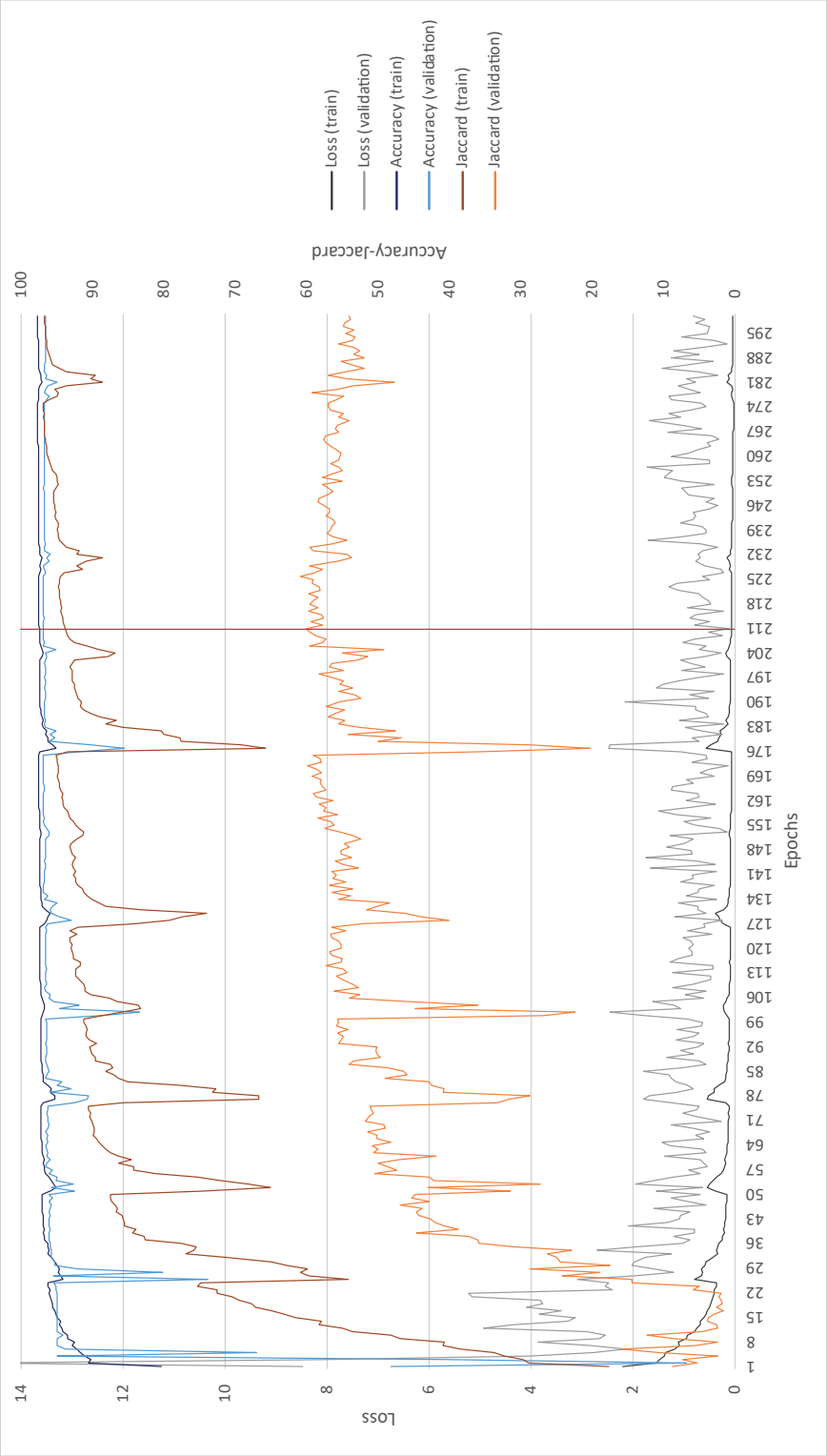
Loss rate, accuracy and Jaccard during training for the train and validation sets using the focal Tversky loss function. The red line indicates the best epoch.



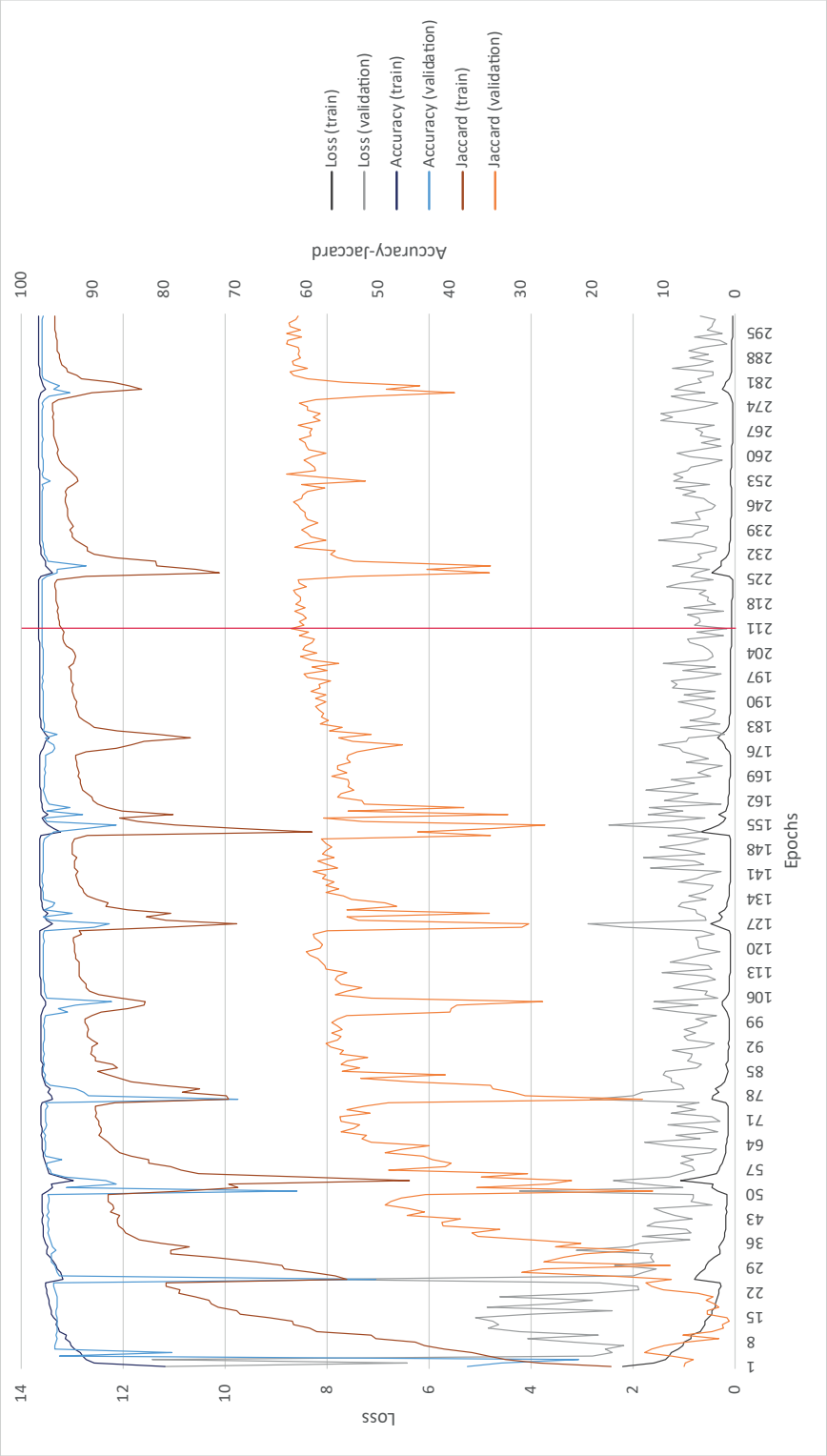
Loss rate, accuracy and Jaccard during training for the train and validation sets using the Lovász-Hinge loss function. The red line indicates the best epoch.



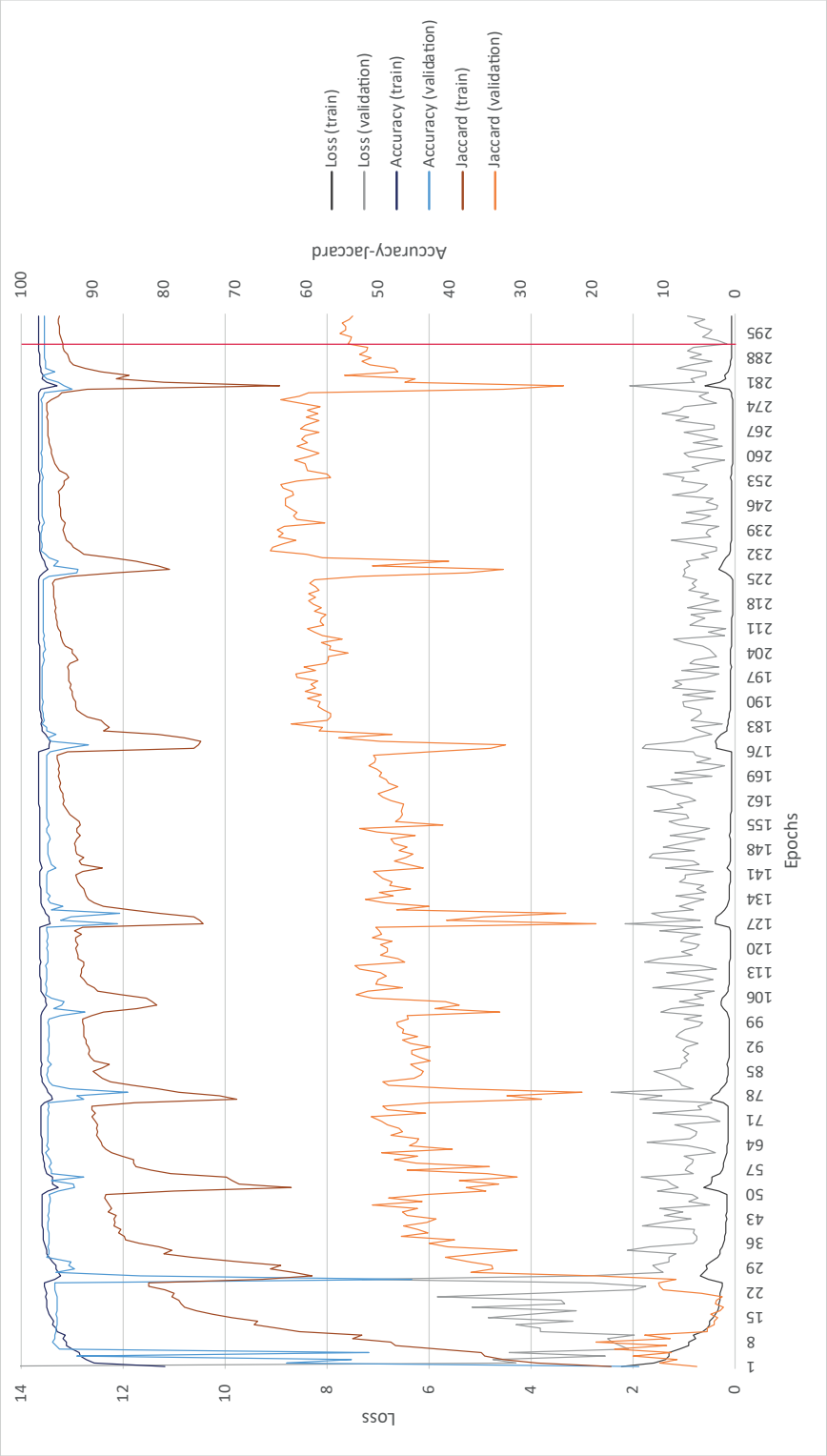
Loss rate, accuracy and Jaccard during training for the train and validation sets using the sum loss function



Loss rate, accuracy and Jaccard during training for the train and validation sets using the mean loss function. The red line indicates the best epoch.

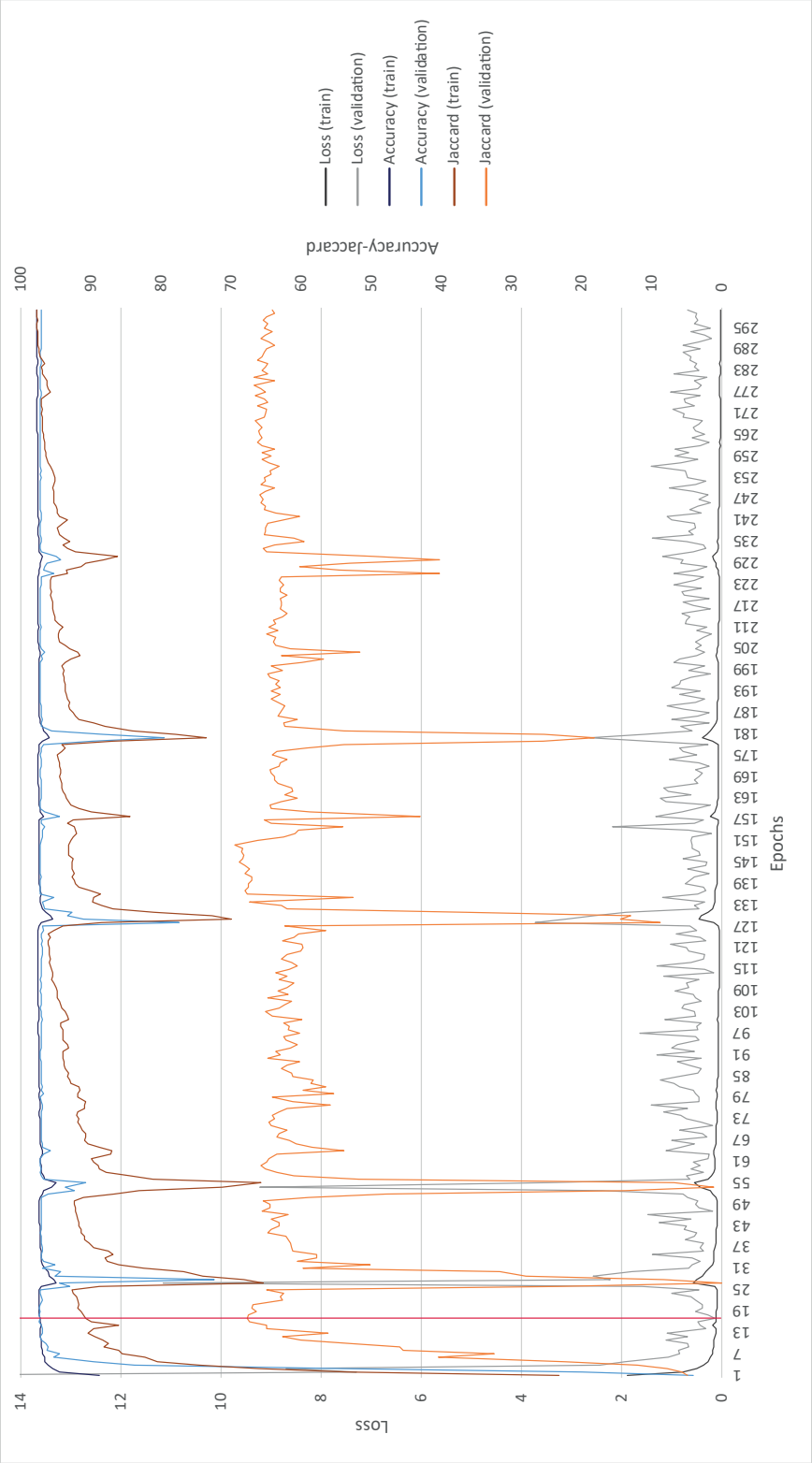


Loss rate, accuracy and Jaccard during training for the train and validation sets using the eigenloss function.
The red line indicates the best epoch.

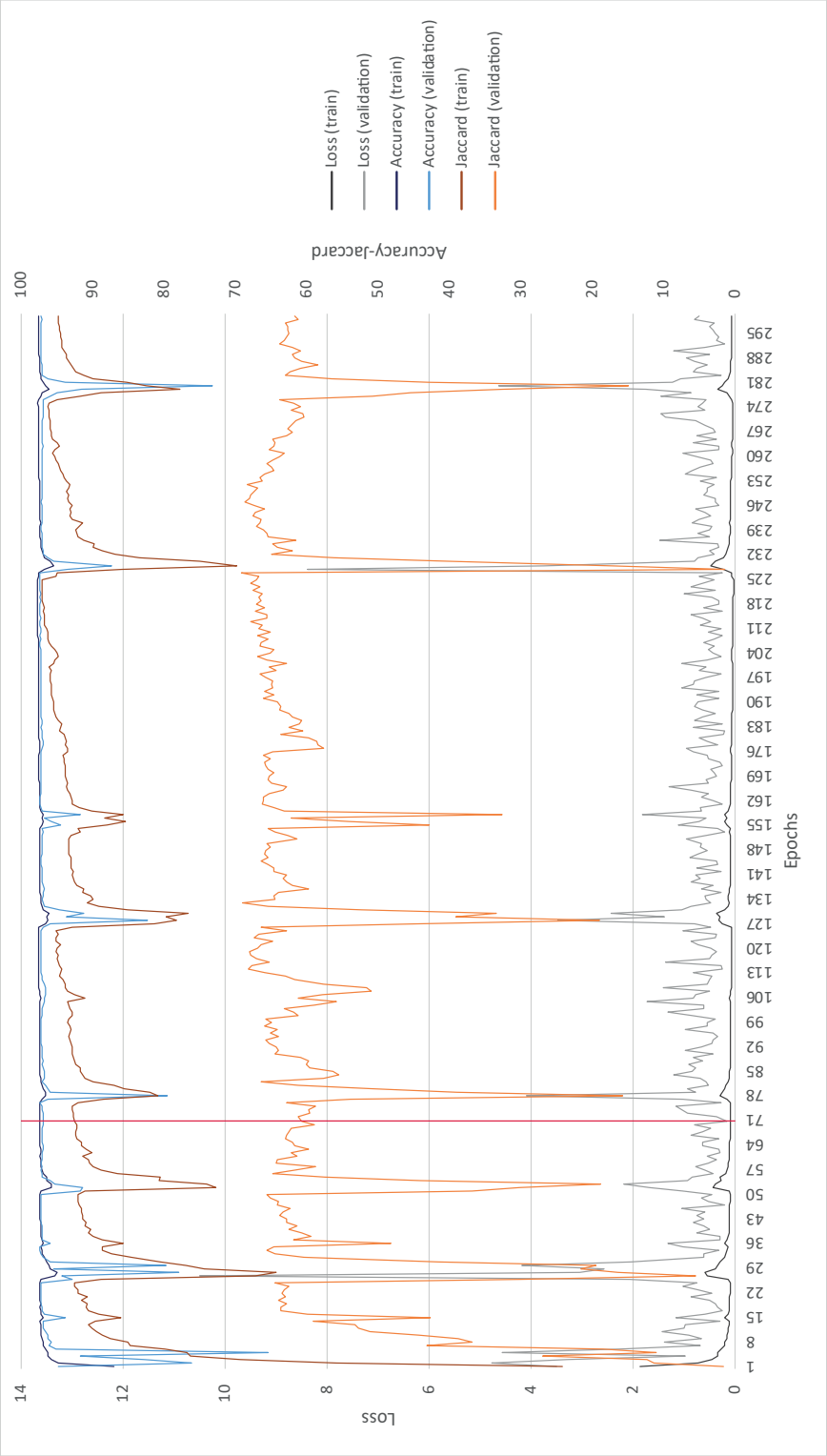


Loss rate, accuracy and Jaccard during training for the train and validation sets using the normalized eigenloss function. The red line indicates the best epoch.

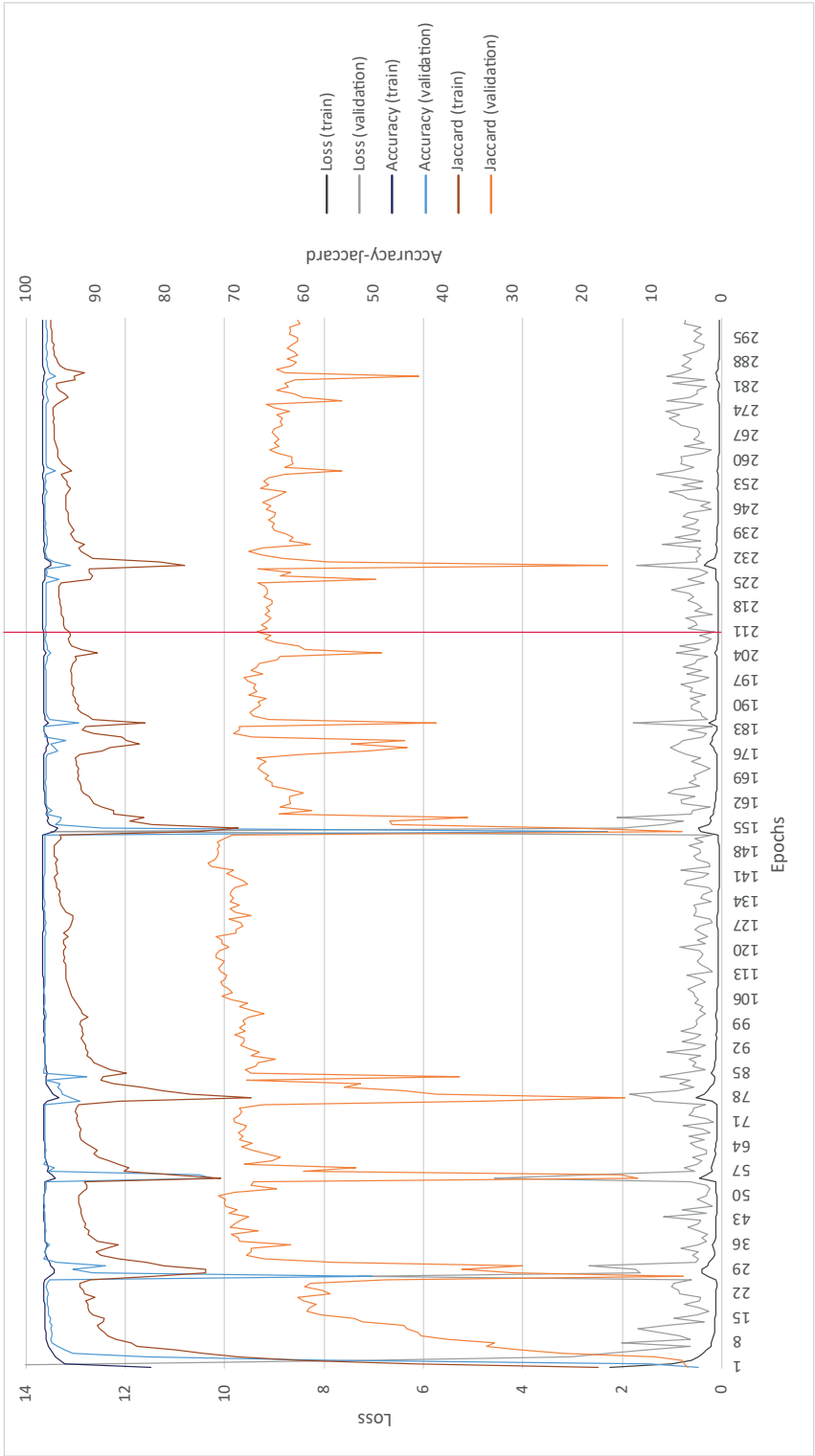
A.3.17 Pérdidas, exactitud e índice de Jaccard durante el entrenamiento para los conjuntos de entrenamiento y validación de LinkNet-Densenet121



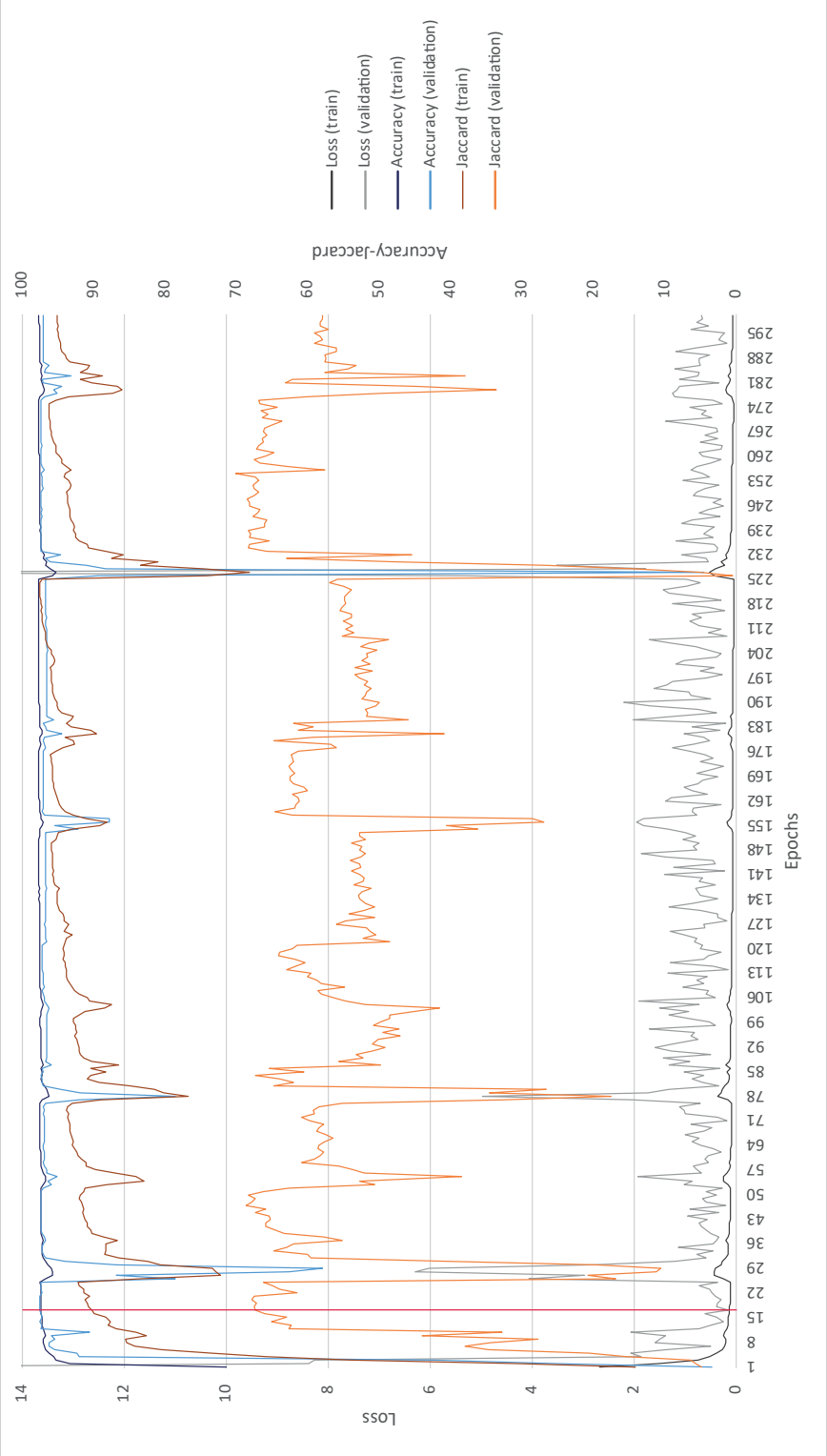
Loss rate, accuracy and Jaccard during training for the train and validation sets using the Jaccard loss function.
The red line indicates the best epoch.



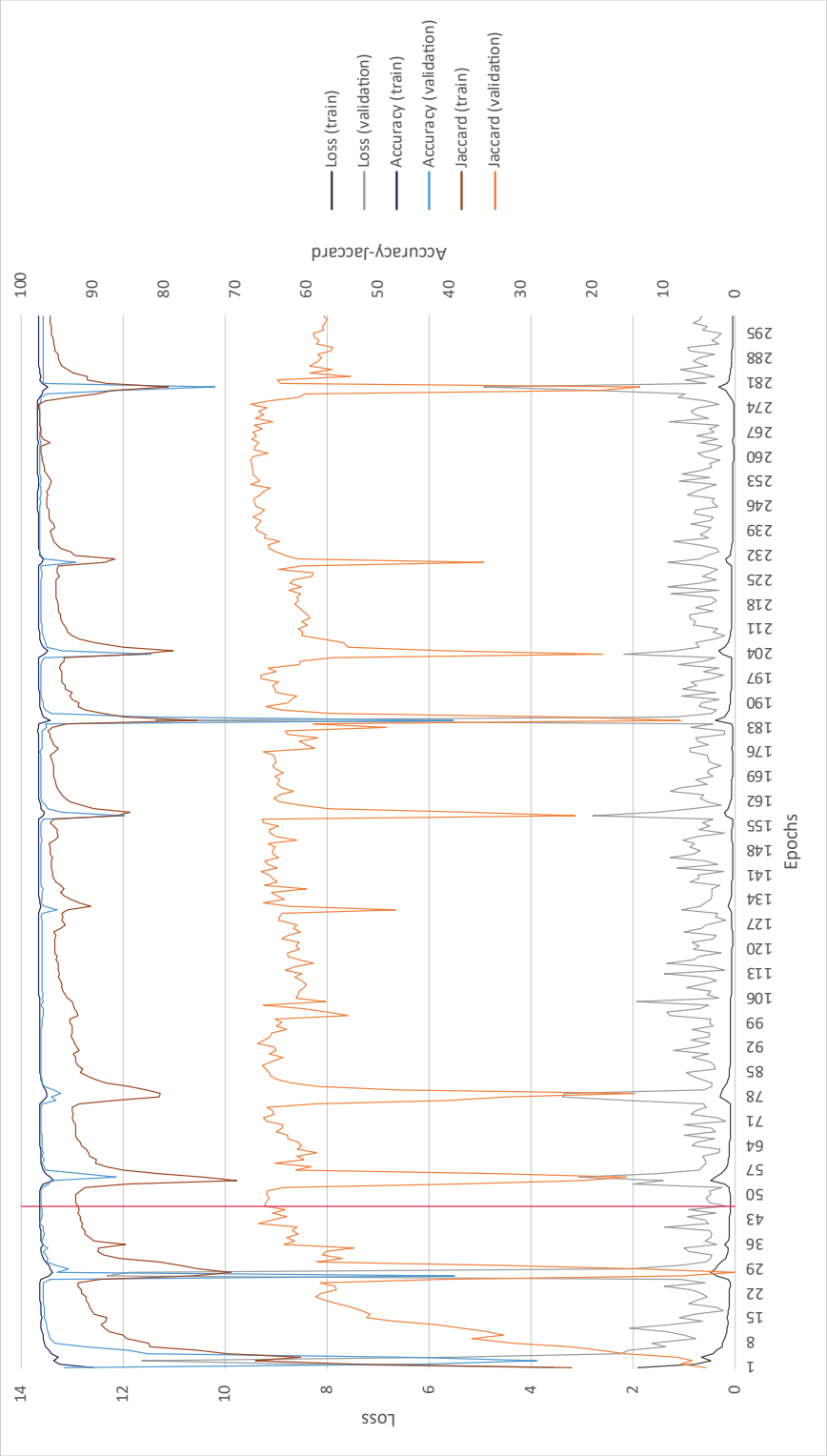
Loss rate, accuracy and Jaccard during training for the train and validation sets using the Dice loss function.
The red line indicates the best epoch.



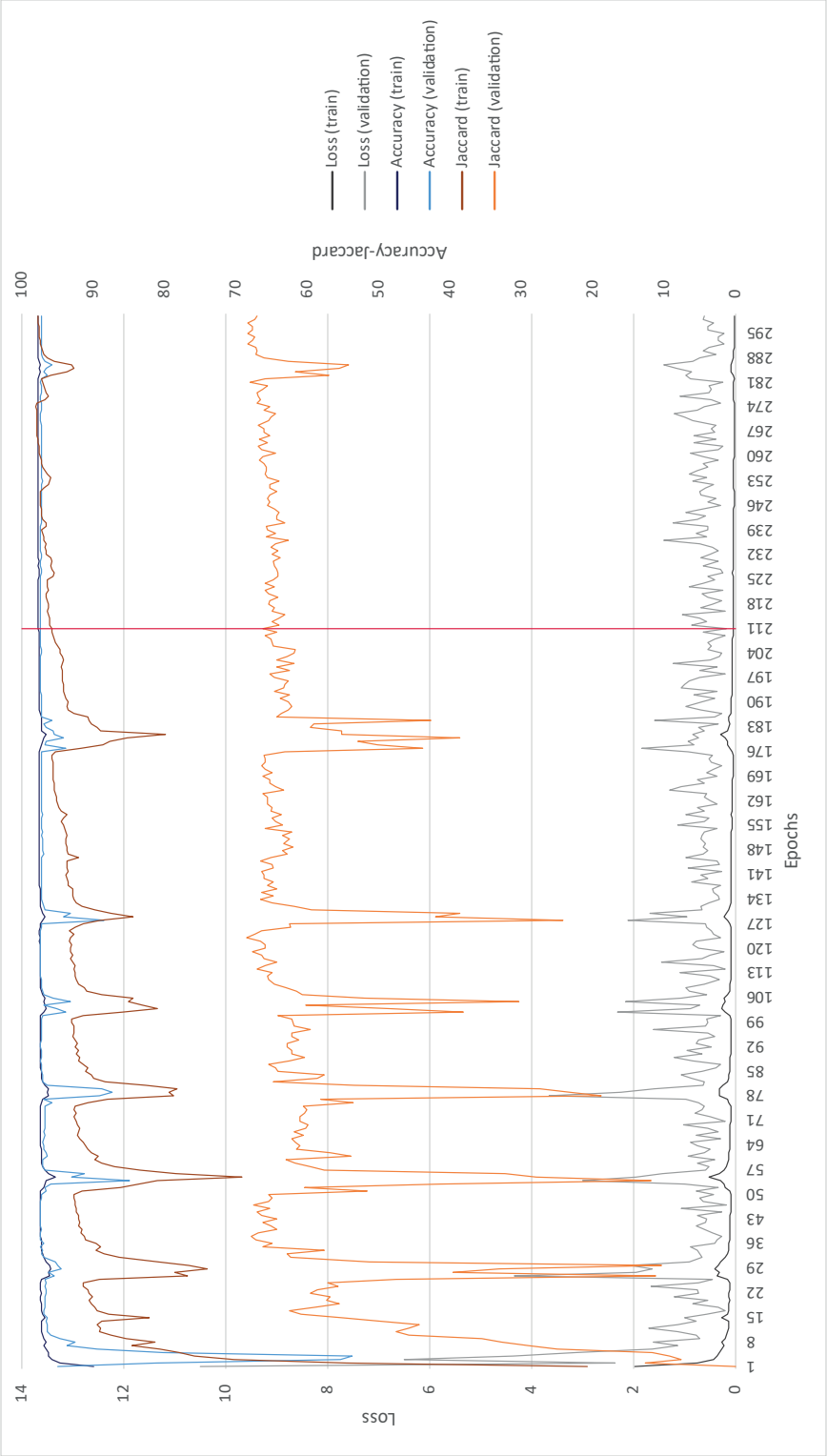
Loss rate, accuracy and Jaccard during training for the train and validation sets using the binary cross entropy loss function. The red line indicates the best epoch.



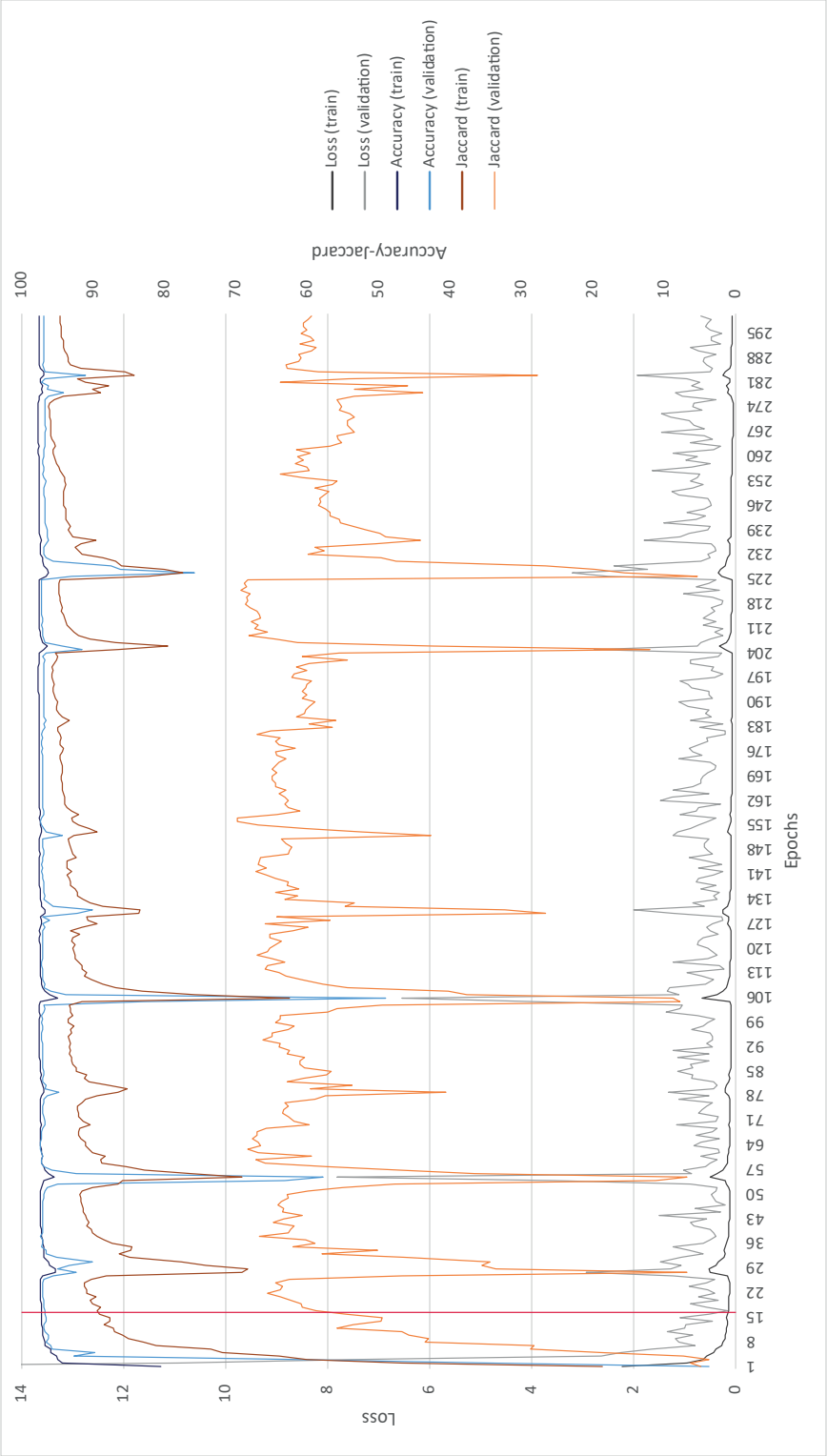
Loss rate, accuracy and Jaccard during training for the train and validation sets using the binary focal loss function. The red line indicates the best epoch.



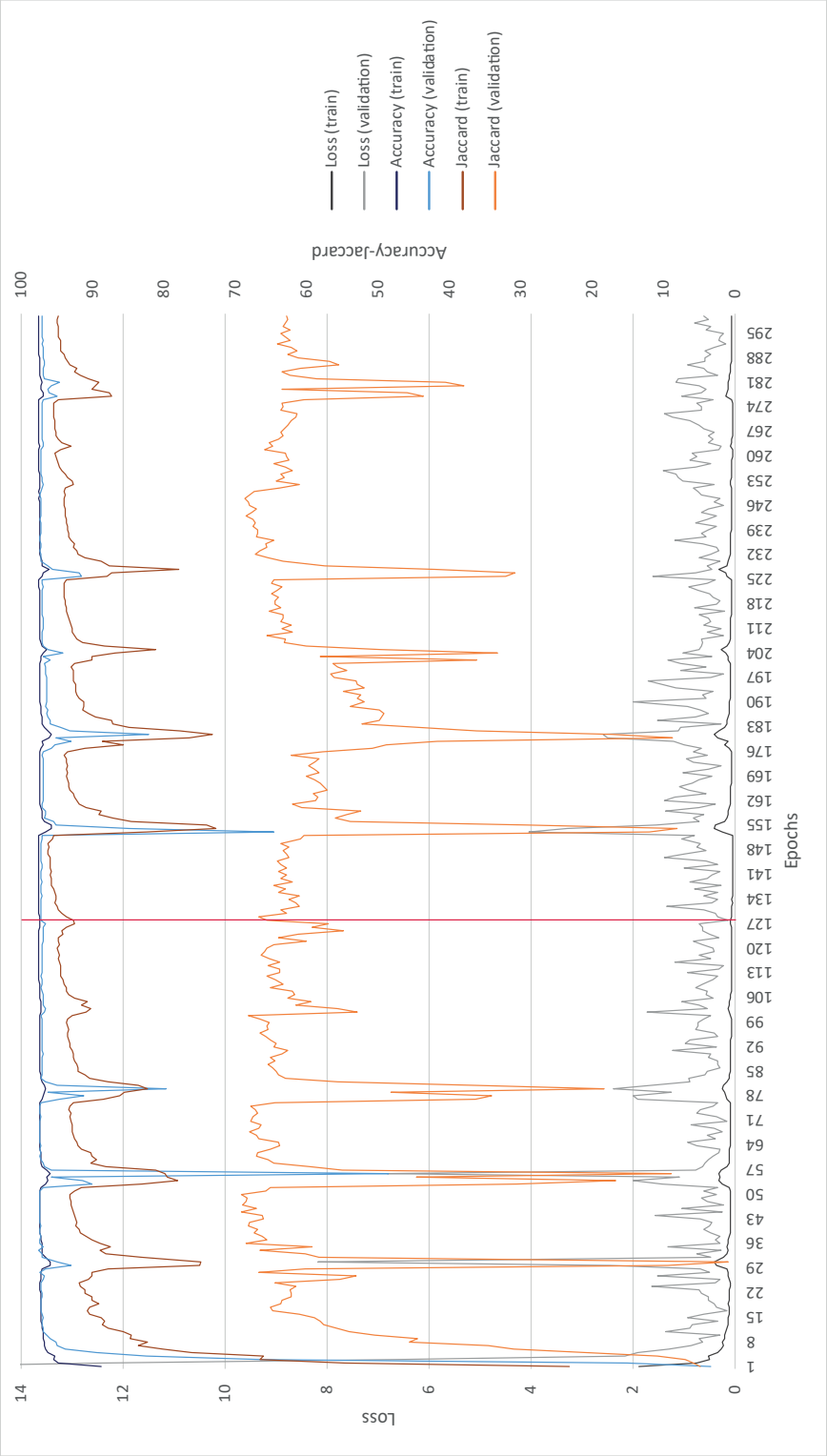
Loss rate, accuracy and Jaccard during training for the train and validation sets using the Tversky loss function.
The red line indicates the best epoch.



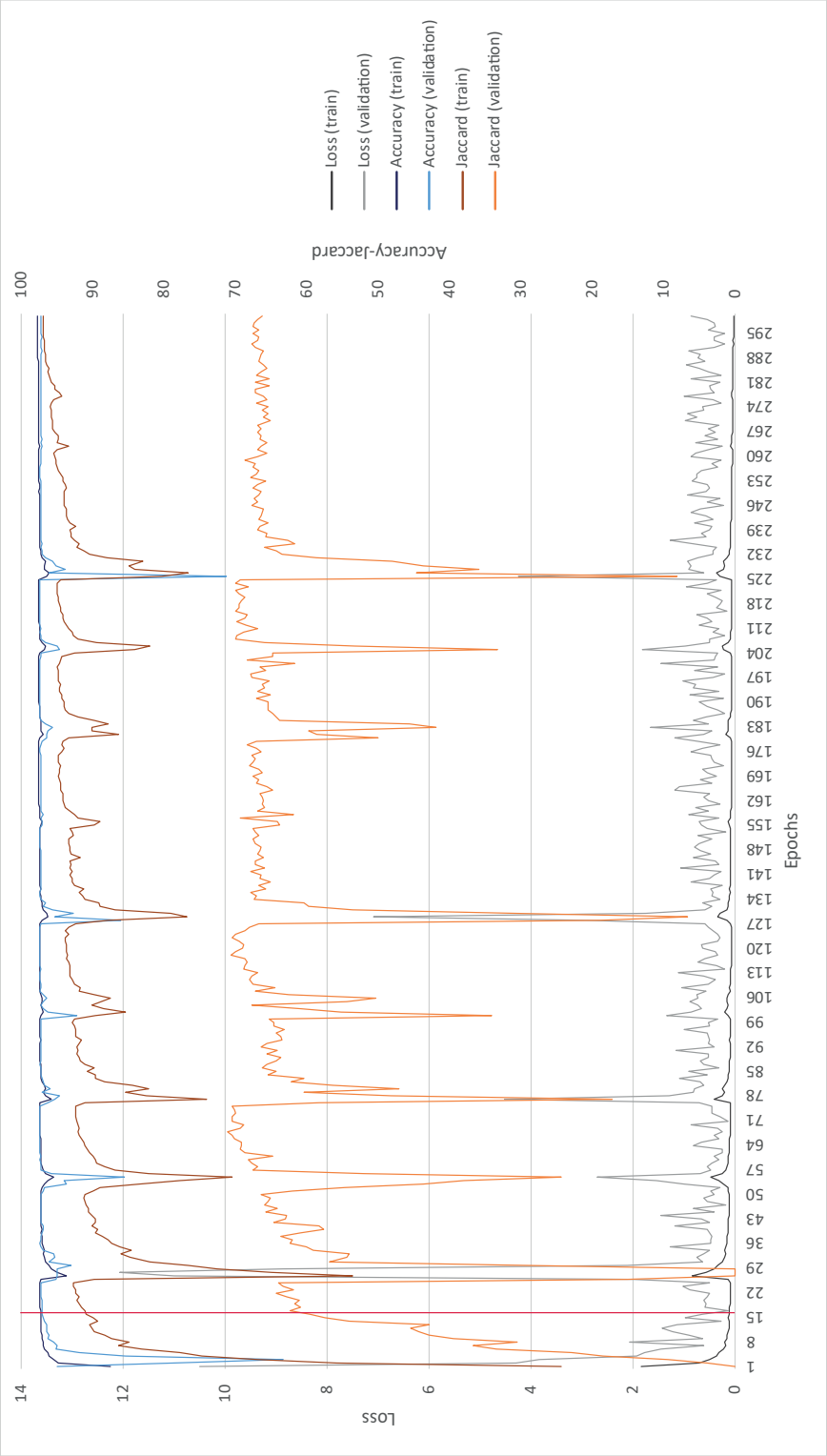
Loss rate, accuracy and Jaccard during training for the train and validation sets using the focal Tversky loss function. The red line indicates the best epoch.



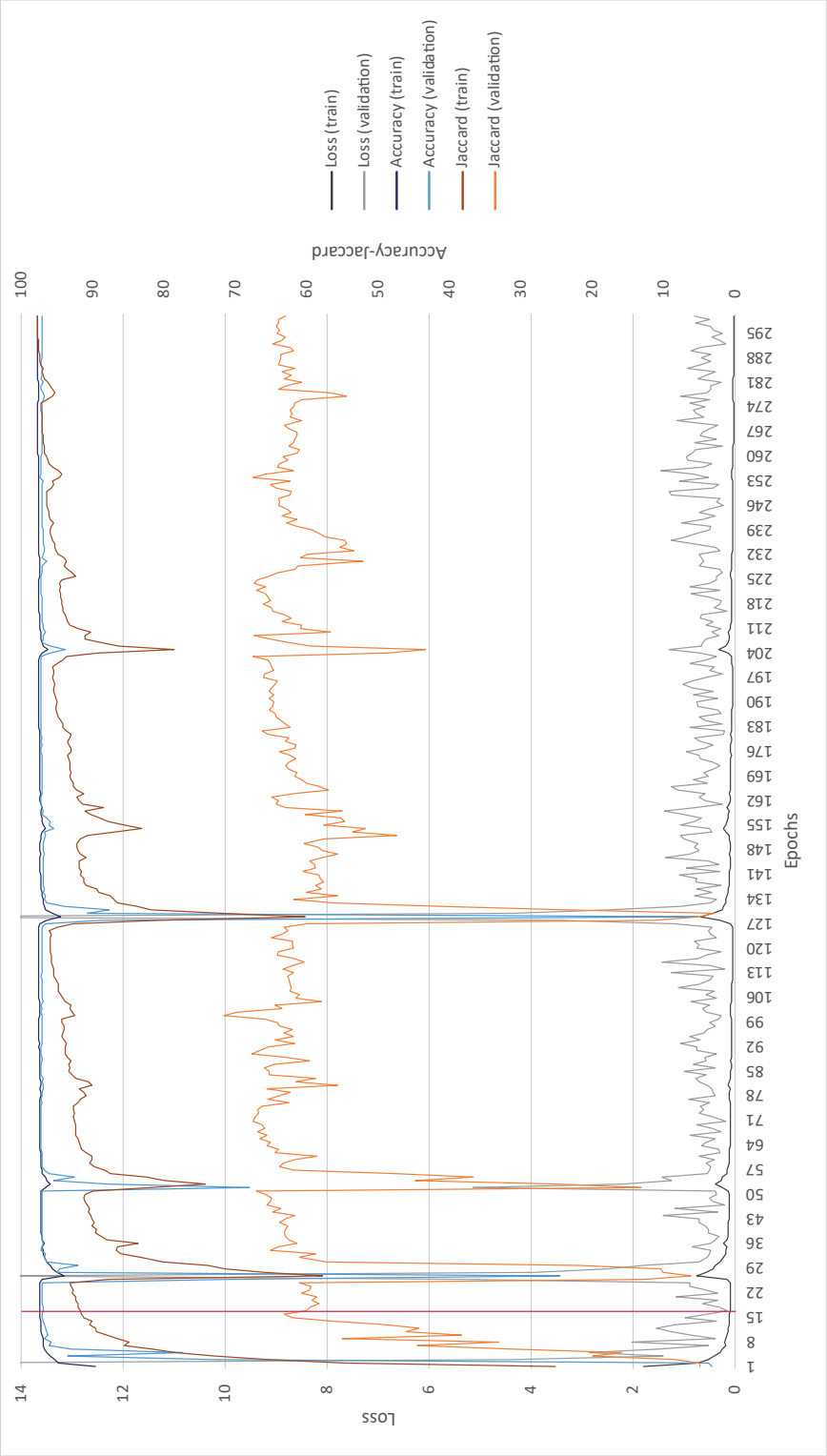
Loss rate, accuracy and Jaccard during training for the train and validation sets using the Lovász-Hinge loss function. The red line indicates the best epoch.



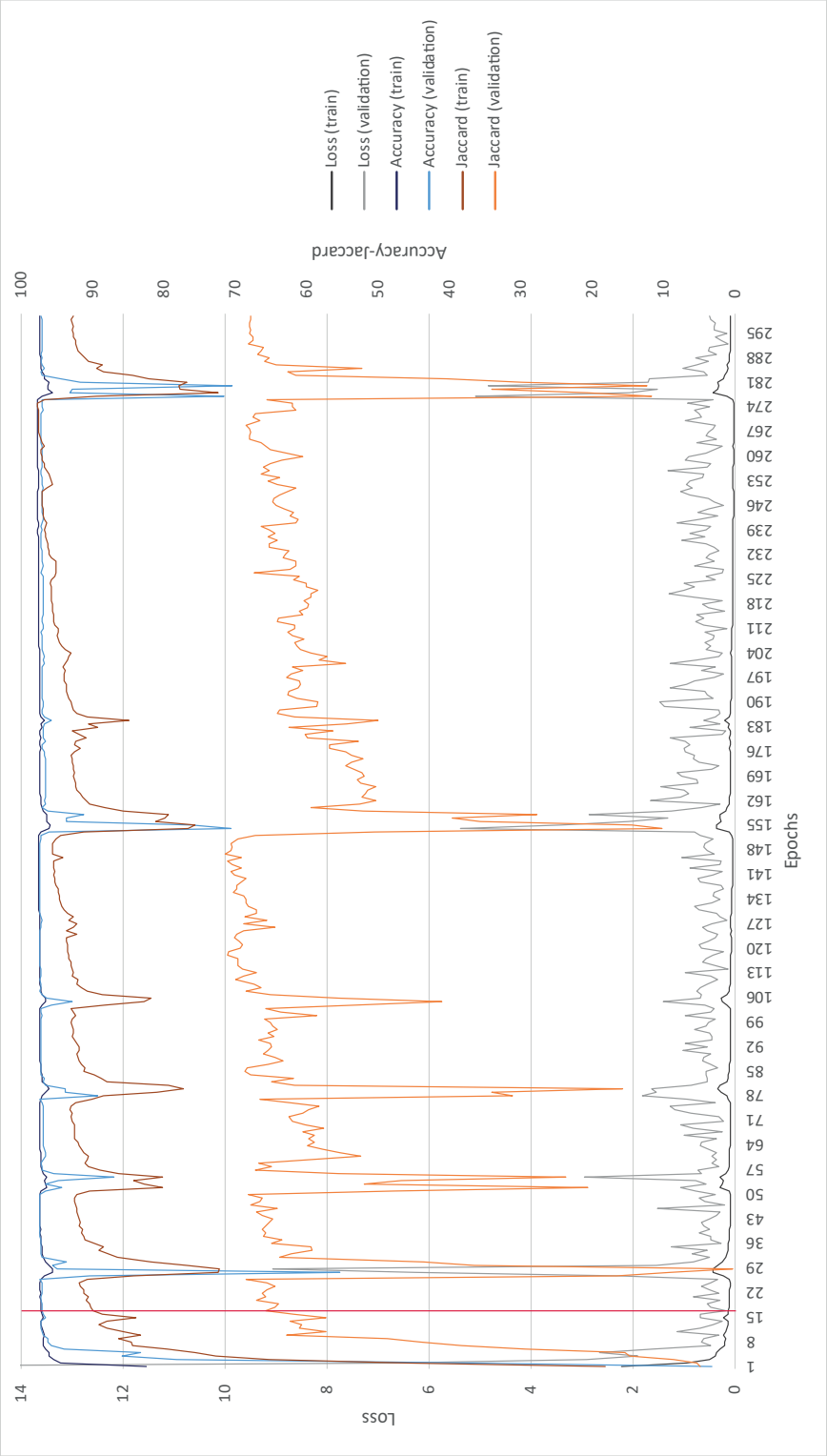
Loss rate, accuracy and Jaccard during training for the train and validation sets using the sum loss function.
The red line indicates the best epoch.



Loss rate, accuracy and Jaccard during training for the train and validation sets using the mean loss function.
The red line indicates the best epoch.



Loss rate, accuracy and Jaccard during training for the train and validation sets using the eigenloss function.
The red line indicates the best epoch.



Loss rate, accuracy and Jaccard during training for the train and validation sets using the normalized eigenloss function. The red line indicates the best epoch.

A.4 PICCOLO white-light and narrow-band imaging colonoscopic dataset: a performance comparative of models and datasets

El material suplementario del artículo *PICCOLO white-light and narrow-band imaging colonoscopic dataset: a performance comparative of models and datasets* está compuesto de los siguientes documentos:

A.4.1 - Metadatos clínicos y distribución de las lesiones en conjuntos de entrenamiento, validación y testeo
A.4.2 - Métricas obtenidas en el conjunto de testeo CVC-EndoSceneStill para los diferentes modelos y conjuntos de datos utilizados durante el entrenamiento
A.4.3 - Métricas obtenidas en el conjunto de testeo Kvasir-SEG para los diferentes modelos y conjuntos de datos utilizados durante el entrenamiento
A.4.4 - Métricas obtenidas en el conjunto de testeo PICCOLO para los diferentes modelos y conjuntos de datos utilizados durante el entrenamiento
A.4.5 - Métricas obtenidas para todos los conjuntos de testeo en su conjunto para los diferentes modelos y conjuntos de datos utilizados durante el entrenamiento
A.4.6 - Imágenes no informativas

A.4.1 Metadatos clínicos y distribución de las lesiones en conjuntos de entrenamiento, validación y testeo

ID Lesión	Video	Núm. pólipos	Pólipo actual	Tamaño (mm)	Clasificación de París	Clasificación de NICE	Diagnóstico preliminar	Diagnóstico literal	Diagnóstico histológico	Subconjunto
1	VP1	1	(1/1)	3	0-lla	Tipo 2	Adenoma	Adenoma	No displasia	Entrenamiento
2	VP2	1	(1/1)	5	Ip	Tipo 2	Adenoma	Adenoma	No displasia	Validación
3	VP3	1	(1/1)	7	Ips	Tipo 2	Adenoma	Adenoma	No displasia	Testeo
4	VP4	2	(1/2)	2	0-lla/c	Tipo 2	Adenoma	Adenoma	No displasia	Testeo
5	VP4	2	(2/2)	3	0-lla	Tipo 2	Adenoma	Adenoma	No displasia	Testeo
6	VP5	2	(1/2)	3	0-lla	Tipo 2	Adenoma	Adenoma	No displasia	Entrenamiento
7	VP5	2	(2/2)	12	Ip	Tipo 2	Adenoma	Adenoma	No displasia	Entrenamiento
8	VP5	2	(2/2)	12	Ip	Tipo 2	Adenoma	Adenoma	No displasia	Entrenamiento
9	VP7	1	(1/1)	3	Is	Tipo 1	Hiperplasia	Hiperplasia	Hiperplasia	Testeo
10	VP9	2	(1/2)	7	Ip	Tipo 2	Adenoma	Adenoma	Displasia de bajo grado	Testeo
11	VP9	2	(2/2)	-	-	Tipo 3	Adenocarcinoma	Adenocarcinoma	Invasive adenocarcinoma	Testeo
12	VP9	2	(2/2)	-	-	Tipo 3	Adenocarcinoma	Adenocarcinoma	Invasive adenocarcinoma	Testeo
13	VP10	1	(1/1)	5	0-lla	Tipo 2	Adenoma	Adenoma	No displasia	Entrenamiento
14	VP12	2	(1/2)	2	0-lla	Tipo 1	Hiperplasia	Hiperplasia	Hiperplasia	Testeo
15	VP12	2	(2/2)	4	Is	Tipo 2	Adenoma	Adenoma	No displasia	Testeo
16	VP13	1	(1/1)	12	Ips	Tipo 2	Adenoma	Adenoma	Displasia de alto grado	Entrenamiento
17	VP14	2	(1/2)	2	0-lla	Tipo 1	Hiperplasia	Hiperplasia	Hiperplasia	Testeo
18	VP14	2	(2/2)	-	-	Tipo 3	Adenocarcinoma	Adenocarcinoma	Invasive adenocarcinoma	Testeo
19	VP22	2	(1/2)	12	Ip	Tipo 2	Adenoma	Adenoma	No displasia	Entrenamiento
20	VP22	2	(2/2)	4	Is	Tipo 2	Adenoma	Adenoma	No displasia	Entrenamiento
21	VP39	2	(1/2)	20	Ips	Tipo 2	Adenoma	Adenoma	Displasia de alto grado	Validación
22	VP39	2	(2/2)	20	0-lla/c	Tipo 2	Adenoma	Adenoma	No displasia	Validación
23	VP40	3	(1/3)	23	0-lla	Tipo 1	Hiperplasia	Hiperplasia	Hiperplasia	Entrenamiento
24	VP40	3	(2/3)	13	0-lla	Tipo 1	Hiperplasia	Hiperplasia	Hiperplasia	Entrenamiento
25	VP40	3	(3/3)	8	0-lla	Tipo 1	Hiperplasia	Hiperplasia	Hiperplasia	Entrenamiento
26	VP41	1	(1/1)	20	0-lla	Tipo 2	Adenoma	Adenoma	Displasia de alto grado	Entrenamiento
27	VP42	1	(1/1)	10	Is	Tipo 2	Adenoma	Adenoma	No displasia	Entrenamiento
28	VP44	2	(1/2)	-	-	Tipo 3	Adenocarcinoma	Adenocarcinoma	Invasive adenocarcinoma	Entrenamiento
29	VP44	2	(2/2)	3	Ips	Tipo 2	Adenoma	Adenoma	No displasia	Entrenamiento
30	VP45	1	(1/1)	-	-	Tipo 3	Adenocarcinoma	Adenocarcinoma	Invasive adenocarcinoma	Validación
31	VP46	1	(1/1)	30	0-lla	Tipo 2	Adenoma	Adenoma	Displasia de alto grado	Entrenamiento
32	VP30	3	(1/3)	6	0-lla	Tipo 2	Adenoma	Adenoma	No displasia	Entrenamiento
33	VP30	3	(2/3)	3	0-lla	Tipo 2	Adenoma	Adenoma	No displasia	Entrenamiento
34	VP30	3	(3/3)	7	0-lla	Tipo 1	Hiperplasia	Hiperplasia	Hiperplasia	Entrenamiento
35	VP31	4	(1/4)	4	0-lla	Tipo 2	Adenoma	Adenoma	No displasia	Validación
36	VP31	4	(2/4)	6	0-lla	Tipo 1	Hiperplasia	Hiperplasia	Hiperplasia	Validación
37	VP31	4	(3/4)	15	0-lla	Tipo 1	Hiperplasia	Hiperplasia	Hiperplasia	Validación
38	VP31	4	(4/4)	10	0-llb	Tipo 1	Hiperplasia	Hiperplasia	Hiperplasia	Validación
39	VP32	1	(1/1)	8	0-lla	Tipo 1	Hiperplasia	Hiperplasia	Hiperplasia	Entrenamiento
40	VP33	2	(1/2)	20	Ips	Tipo 2	Adenoma	Adenoma	No displasia	Entrenamiento
41	VP33	2	(2/2)	50	Ips	Tipo 2	Adenoma	Adenoma	Displasia de alto grado	Entrenamiento
42	VP34	3	(1/3)	12	0-lla	Tipo 1	Hiperplasia	Hiperplasia	Hiperplasia	Entrenamiento
43	VP34	3	(2/3)	6	0-lla	Tipo 1	Hiperplasia	Hiperplasia	Hiperplasia	Entrenamiento
44	VP34	3	(3/3)	15	Ip	Tipo 2	Adenoma	Adenoma	No displasia	Entrenamiento
45	VP35	1	(1/1)	30	Is	Tipo 2	Adenoma	Adenoma	Displasia de alto grado	Validación
46	VP35	1	(1/1)	30	Is	Tipo 2	Adenoma	Adenoma	Displasia de alto grado	Validación
47	VP36	1	(1/1)	50	Is	Tipo 2	Adenoma	Adenoma	Displasia de alto grado	Validación
48	VP36	1	(1/1)	50	Is	Tipo 2	Adenoma	Adenoma	Displasia de alto grado	Validación
49	VP37	2	(1/2)	32	Is					Entrenamiento
50	VP37	2	(2/2)	9	Ip	Tipo 2	Adenoma	Adenoma	No displasia	Entrenamiento
51	VP38	2	(1/2)	10	0-lla	Tipo 2	Adenoma	Adenoma	No displasia	Entrenamiento
52	VP38	2	(2/2)	15	Ip	Tipo 2	Adenoma	Adenoma	No displasia	Entrenamiento
53	VP47	1	(1/1)	-	-	Tipo 3	Adenocarcinoma	Adenocarcinoma	Invasive adenocarcinoma	Validación
54	VP48	2	(1/2)	30	Is	Tipo 2	Adenoma	Adenoma	Displasia de alto grado	Entrenamiento
55	VP48	2	(2/2)	-	-	Tipo 3	Adenocarcinoma	Adenocarcinoma	Invasive adenocarcinoma	Entrenamiento
56	VP48	2	(2/2)	-	-	Tipo 3	Adenocarcinoma	Adenocarcinoma	Invasive adenocarcinoma	Entrenamiento
57	VP15	2	(1/2)	10	0-lla	Tipo 1	Hiperplasia	Hiperplasia	No displasia	Entrenamiento
58	VP15	2	(2/2)	8	Is	Tipo 2	Adenoma	Adenoma	No displasia	Entrenamiento
59	VP17	2	(1/2)	28	0-lla	Tipo 2	Adenoma	Adenoma	No displasia	Validación
60	VP17	2	(2/2)	31	0-lla	Tipo 2	Adenoma	Adenoma	No displasia	Validación
61	VP18	2	(1/2)	5	0-lla	Tipo 2	Adenoma	Adenoma	No displasia	Validación
62	VP18	2	(2/2)	8	Ip	Tipo 2	Adenoma	Adenoma	No displasia	Validación
63	VP19	1	(1/1)	5	Is	Tipo 2	Adenoma	Adenoma	No displasia	Entrenamiento
64	VP24	1	(1/1)	4	0-llb	Tipo 1	Hiperplasia	Hiperplasia	No displasia	Testeo
65	VP26	2	(1/2)	7	0-lla	Tipo 2	Adenoma	Adenoma	No displasia	Entrenamiento
66	VP26	2	(2/2)	5	Is	Tipo 2	Adenoma	Adenoma	No displasia	Entrenamiento
67	VP27	1	(1/1)	25	0-lla/c	Tipo 2	Adenoma	Adenoma	Displasia de alto grado	Entrenamiento
68	VP27	1	(1/1)	25	0-lla/c	Tipo 2	Adenoma	Adenoma	Displasia de alto grado	Entrenamiento
69	VP28	1	(1/1)	7	0-lla	Tipo 2	Adenoma	Adenoma	No displasia	Entrenamiento
70	VP29	2	(1/2)	5	0-lla	Tipo 1	Hiperplasia	Hiperplasia	No displasia	Entrenamiento
71	VP29	2	(2/2)	5	0-lla	Tipo 1	Hiperplasia	Hiperplasia	No displasia	Entrenamiento
72	VP21	5	(1/5)	3	Is	Tipo 2	Adenoma	Adenoma	No displasia	Entrenamiento
73	VP21	5	(2/5)	8	Is	Tipo 2	Adenoma	Adenoma	No displasia	Entrenamiento
74	VP21	5	(3/5)	6	Is	Tipo 2	Adenoma	Adenoma	No displasia	Entrenamiento
75	VP21	5	(4/5)	9	Ips	Tipo 2	Adenoma	Adenoma	No displasia	Entrenamiento
76	VP21	5	(5/5)	10	Ip	Tipo 2	Adenoma	Adenoma	No displasia	Entrenamiento

A.4.2 Métricas obtenidas en el conjunto de testeo CVC-EndoSceneStill para los diferentes modelos y conjuntos de datos utilizados durante el entrenamiento

Table S2. Metrics obtained in the CVC-EndoSceneStill test set for the different models and datasets used during training. Best value per metric is indicated in bold.

Dataset for training/validation set	Network	CVC-EndoSceneStill						
		Accuracy	Precision	Recall	Specificity	F2-score	Jaccard	Dice
CVC-EndoSceneStill	U-Net+VGG16	92.77±10.50	77.84±34.17	63.02±39.00	99.04±2.12	62.98±37.91	56.80±36.09	64.11±36.83
	U-Net+Densenet121	94.10±10.53	**88.87±22.15**	72.46±36.84	**99.38±1.55**	72.59±35.67	67.31±34.47	73.63±33.60
	Linknet+VGG16	93.58±10.24	85.05±26.50	69.40±36.28	99.36±1.25	69.23±34.84	62.77±33.81	70.39±33.03
	Linknet+Densenet121	94.83±9.60	87.88±21.73	79.22±32.39	99.00±2.15	78.54±31.19	**72.16±30.93**	78.61±29.48
Kvasir-SEG	U-Net+VGG16	91.76±10.36	55.43±47.53	34.56±40.81	99.36±2.63	35.15±40.66	32.44±38.79	37.11±41.12
	U-Net+Densenet121	94.74±9.94	83.54±25.84	**83.46±28.83**	97.65±7.17	**80.10±28.71**	71.82±29.87	**78.78±28.14**
	Linknet+VGG16	93.71±9.62	79.03±33.46	64.45±39.17	99.19±1.68	64.52±38.03	58.62±36.18	65.69±36.70
	Linknet+Densenet121	**95.32±9.07**	82.98±27.51	76.72±34.88	98.77±3.83	75.79±33.99	69.10±32.53	75.55±32.58
PICCOLO	U-Net+VGG16	91.03±10.99	66.18±38.07	56.96±40.54	96.91±6.10	55.73±39.19	47.76±36.46	55.43±37.98
	U-Net+Densenet121	93.86±9.97	79.31±33.02	68.99±36.56	99.03±2.45	68.76±35.75	62.41±34.78	69.46±35.13
	Linknet+VGG16	92.78±10.45	78.42±32.36	61.14±38.17	98.74±3.02	60.99±37.00	54.58±35.63	62.46±36.01
	Linknet+Densenet121	93.77±10.24	84.35±25.26	71.88±34.26	98.85±2.86	71.05±33.40	64.18±33.04	71.66±32.98

A.4.3 Métricas obtenidas en el conjunto de testeo Kvasir-SEG para los diferentes modelos y conjuntos de datos utilizados durante el entrenamiento

Table S3. Metrics obtained in the Kvasir-SEG test set for the different models and datasets used during training. Best value per metric is indicated in bold.

Dataset for training/validation set	Network	Kvasir-SEG						
		Accuracy	Precision	Recall	Specificity	F2-score	Jaccard	Dice
CVC-EndoSceneStill	U-Net+VGG16	87.63±13.25	67.59±32.87	61.51±33.74	94.68±8.62	57.64±31.54	46.29±29.36	57.24±30.72
	U-Net+Densenet121	90.53±14.05	77.80±32.68	63.31±37.22	98.15±6.23	62.72±36.38	56.12±34.29	64.26±35.35
	Linknet+VGG16	87.42±12.73	65.39±30.27	73.37±30.03	92.44±9.58	66.06±28.55	51.13±27.67	62.61±28.17
	Linknet+Densenet121	90.26±13.90	74.26±34.70	64.89±36.51	**97.99±4.29**	64.02±35.45	56.69±33.68	65.05±34.66
Kvasir-SEG	U-Net+VGG16	93.34±8.66	79.93±27.73	75.04±32.90	97.77±3.85	73.30±31.08	64.23±29.78	72.94±29.76
	U-Net+Densenet121	94.64±7.75	81.03±21.93	88.89±22.25	96.29±5.89	85.36±21.58	74.13±23.40	82.28±21.76
	Linknet+VGG16	94.27±8.12	**82.81±21.49**	85.26±23.07	97.10±4.98	82.62±22.36	72.53±23.92	81.09±22.08
	Linknet+Densenet121	**94.70±7.82**	81.68±21.52	**88.98±21.26**	96.49±5.86	**85.46±20.93**	**74.52±22.81**	**82.68±21.28**
PICCOLO	U-Net+VGG16	88.17±13.14	64.41±31.66	74.74±32.23	91.89±11.77	66.91±30.63	52.64±30.41	62.97±30.74
	U-Net+Densenet121	92.57±11.23	81.31±26.76	74.73±31.82	97.41±4.89	73.43±31.04	65.33±30.66	73.54±30.15
	Linknet+VGG16	91.02±10.94	72.60±30.55	74.73±32.22	95.49±7.08	70.55±30.58	58.99±30.40	68.54±30.28
	Linknet+Densenet121	91.29±13.63	75.91±31.17	70.37±35.96	95.74±11.03	68.09±34.81	59.61±33.80	67.58±34.14

A.4.4 Métricas obtenidas en el conjunto de testeo PICCOLO para los diferentes modelos y conjuntos de datos utilizados durante el entrenamiento

Table S4. Metrics obtained in the PICCOLO test set for the different models and datasets used during training. Best value per metric is indicated in bold.

Dataset for training/validation set	Network	PICCOLO – WL and NBI frames						
		Accuracy	Precision	Recall	Specificity	F2-score	Jaccard	Dice
CVC-EndoSceneStill	U-Net+VGG16	81.97±19.67	55.84±44.84	40.59±39.2	94.19±12.64	38.06±37.81	32.82±34.81	39.39±38.72
	U-Net+Densenet121	84.65±19.63	55.31±45.60	42.92±42.12	97.26±6.89	42.66±41.85	38.76±39.05	44.10±42.21
	Linknet+VGG16	83.81±19.04	54.78±44.67	34.60±37.84	97.82±4.09	35.21±37.34	30.55±33.57	37.23±37.73
	Linknet+Densenet121	85.32±19.01	52.84±45.12	45.09±42.36	97.30±6.26	44.78±41.60	39.52±37.9	45.5±41.51
Kvasir-SEG	U-Net+VGG16	84.70±19.09	52.00±46.99	30.86±38.27	**99.41±2.16**	31.74±38.24	28.51±35.12	33.89±38.82
	U-Net+Densenet121	87.76±17.47	60.77±43.16	53.77±43.26	96.97±8.57	51.38±41.52	44.78±38.73	51.00±41.08
	Linknet+VGG16	88.31±14.05	63.70±43.49	44.95±40.68	99.01±2.32	45.42±39.81	40.43±36.41	47.43±39.69
	Linknet+Densenet121	86.14±18.45	59.54±42.34	58.11±42.49	93.29±16.63	52.87±40.07	44.92±37.37	51.87±39.79
PICCOLO	U-Net+VGG16	92.66±10.39	77.18±35.21	64.76±38.61	98.50±3.70	64.66±37.67	58.74±36.06	65.74±36.93
	U-Net+Densenet121	92.70±11.61	**77.58±33.87**	**72.89±37.04**	97.96±4.99	**70.69±36.38**	**64.01±36.23**	**70.10±36.45**
	Linknet+VGG16	91.91±11.88	73.16±37.77	60.69±41.91	98.29±4.96	60.02±41.02	54.46±38.88	60.48±40.36
	Linknet+Densenet121	**93.05±11.65**	75.85±36.49	65.36±40.63	98.01±6.30	65.16±39.91	60.14±38.31	65.81±39.32
		PICCOLO – Only WL frames						
CVC-EndoSceneStill	U-Net+VGG16	85.38±14.42	60.93±42.72	57.06±37.65	91.12±14.91	52.02±37.24	45.00±35.60	52.81±38.33
	U-Net+Densenet121	89.53±10.67	62.86±44.81	50.27±40.71	96.87±5.97	49.73±40.66	45.57±38.71	51.78±41.27
	Linknet+VGG16	88.36±11.47	59.47±43.55	42.08±37.54	97.10±4.21	42.81±36.80	37.17±33.45	45.31±37.13
	Linknet+Densenet121	90.30±9.47	57.08±45.26	50.55±41.30	96.74±5.02	50.28±40.96	45.09±38.42	51.33±41.47
Kvasir-SEG	U-Net+VGG16	88.59±13.40	42.90±47.74	30.56±38.28	**99.43±2.62**	31.61±38.86	29.01±36.31	33.77±40.32
	U-Net+Densenet121	92.36±10.67	63.61±43.66	54.63±42.64	98.06±6.63	53.13±41.67	47.74±39.55	53.62±41.68
	Linknet+VGG16	89.78±12.39	58.05±46.59	36.46±39.19	99.57±1.35	37.71±39.50	34.53±36.88	40.39±40.52
	Linknet+Densenet121	92.40±9.86	63.26±44.32	49.02±42.33	99.19±1.52	49.56±41.94	45.18±39.01	51.13±41.90
PICCOLO	U-Net+VGG16	93.21±10.81	**73.52±39.73**	57.79±41.84	99.09±3.51	58.32±41.58	54.38±40.08	59.83±41.42
	U-Net+Densenet121	92.86±11.11	71.33±38.15	**69.77±39.53**	97.37±6.29	**65.97±38.64**	**58.70±38.90**	**64.51±39.18**
	Linknet+VGG16	92.66±11.15	68.43±42.15	54.75±43.72	98.73±4.95	54.79±43.29	50.66±41.13	55.57±43.24
	Linknet+Densenet121	**93.34±11.36**	69.18±40.96	59.12±42.68	98.17±6.62	58.88±42.08	54.47±40.60	59.64±41.94

A.4.5 Métricas obtenidas para todos los conjuntos de testeo en su conjunto para los diferentes modelos y conjuntos de datos utilizados durante el entrenamiento

Table S5. Metrics obtained for all test sets as a whole for the different models and datasets used during training. Best value per metric is indicated in bold.

Dataset for training/validation set	Network	All datasets						
		Accuracy	Precision	Recall	Specificity	F2-score	Jaccard	Dice
CVC-EndoSceneStill	U-Net+VGG16	86.30±16.63	64.72±40.19	52.15±39.17	95.56±10.02	49.88±37.84	42.69±35.13	50.68±37.70
	U-Net+Densenet121	88.70±16.68	70.14±39.97	56.15±41.46	98.05±5.85	55.89±40.88	50.88±38.51	57.26±40.32
	Linknet+VGG16	87.31±16.00	65.46±38.89	54.3±39.900	96.71±6.42	52.50±38.02	44.51±34.86	52.77±37.13
	Linknet+Densenet121	89.12±16.15	67.75±40.26	59.31±40.94	97.93±5.00	58.75±40.00	52.63±37.53	59.39±39.36
Kvasir-SEG	U-Net+VGG16	88.91±15.28	60.69±44.27	44.16±42.17	**98.94±2.93**	44.23±41.22	39.50±37.98	45.63±40.84
	U-Net+Densenet121	91.46±13.99	72.23±35.85	71.15±38.56	96.96±7.56	68.20±37.29	59.87±35.72	66.82±36.57
	Linknet+VGG16	91.35±11.91	72.95±36.99	61.19±39.94	98.52±3.31	60.69±38.60	54.04±35.99	61.49±37.60
	Linknet+Densenet121	90.87±14.69	71.70±35.83	71.48±38.08	95.58±12.13	67.82±37.00	59.35±35.33	66.52±36.41
PICCOLO	U-Net+VGG16	90.99±11.51	70.81±35.48	65.56±37.97	96.25±7.89	63.02±36.47	54.24±34.93	62.34±35.79
	U-Net+Densenet121	**92.96±11.10**	**79.06±31.81**	**72.41±35.54**	98.08±4.48	**70.97±34.80**	**63.97±34.35**	**70.90±34.45**
	Linknet+VGG16	91.88±11.28	74.34±34.56	64.73±38.91	97.62±5.43	63.21±37.56	55.76±35.87	63.24±36.79
	Linknet+Densenet121	92.74±11.94	78.03±32.64	68.42±37.87	97.59±7.47	67.48±36.97	61.02±35.79	67.79±36.40
		All datasets (from PICCOLO, only WL frames are included)						
CVC-EndoSceneStill	U-Net+VGG16	88.51±13.20	68.61±37.41	60.50±36.80	94.87±10.56	57.45±35.79	49.19±34.07	57.94±35.59
	U-Net+Densenet121	91.33±12.05	76.31±36.15	61.85±39.29	98.11±5.21	61.50±38.71	56.14±36.88	63.05±37.91
	Linknet+VGG16	89.69±11.85	69.65±35.90	61.65±37.34	96.19±6.83	59.29±35.44	50.15±33.28	59.29±34.47
	Linknet+Densenet121	91.72±11.43	72.83±37.49	64.64±38.69	97.89±4.13	64.03±37.86	57.72±36.17	64.76±37.23
Kvasir-SEG	U-Net+VGG16	91.25±11.13	59.78±44.54	47.33±42.48	**98.83±3.20**	47.26±41.55	42.37±38.48	48.48±41.30
	U-Net+Densenet121	93.91±9.55	76.00±33.04	**75.71±35.63**	97.32±6.60	72.91±34.68	64.57±33.72	71.59±34.03
	Linknet+VGG16	92.59±10.36	73.33±36.85	62.34±39.88	98.59±3.36	61.86±38.71	55.41±36.29	62.59±37.81
	Linknet+Densenet121	94.13±9.01	75.93±33.70	71.73±37.72	98.12±4.33	**70.39±36.57**	**62.99±34.45**	**69.87±35.54**
PICCOLO	U-Net+VGG16	90.76±11.90	68.02±36.71	63.43±39.14	95.89±8.56	60.49±37.55	51.67±35.83	59.53±36.95
	U-Net+Densenet121	**93.08±10.80**	**77.34±33.12**	71.25±36.07	97.91±4.90	69.45±35.29	62.19±34.94	69.23±35.09
	Linknet+VGG16	92.13±10.87	73.05±35.55	63.74±39.11	97.60±5.55	62.24±37.75	54.80±36.03	62.27±37.15
	Linknet+Densenet121	92.76±11.91	76.34±33.70	67.08±38.20	97.54±7.83	65.95±37.28	59.34±36.14	66.22±36.87

A.4.6 Imágenes no informativas