Politecnico di Torino

PolypGen Challenge

2024-2025

Catania Cecere Eleonora - 327305

Cevoli Elisa - 330775

Cordoba Acosta Tatiana Camila - 320757

Ottomano Claudia - 331014

1 Introduzione

Il carcinoma del colon-retto è una delle principali cause di morte per cancro a livello globale, con 1.93 milioni di nuovi casi registrati nel 2020 [1]. I polipi del colon, in particolare quelli adenomatosi, rappresentano un importante fattore di rischio, essendo lesioni precancerose comuni, con un'incidenza che può superare il 40% nelle persone sopra i 60 anni. La colonscopia è considerata il gold standard per la diagnosi e la rimozione dei polipi [2]. Nonostante i progressi nello screening, il cancro colorettale resta una delle principali cause di mortalità. Questo è spesso dovuto al "cancro colorettale intermedio", causato dalla mancata rimozione di polipi precancerosi durante la colonscopia, anche con sorveglianza regolare [3]. L'intelligenza artificiale migliora la diagnosi rilevando automaticamente anche i polipi più difficili e supportando i gastroenterologi nell'analisi accurata delle lesioni. Questo approccio diminuisce la necessità di ricorrere alla valutazione patologica e agevola decisioni più rapide e precise sulle tecniche di resezione e sui trattamenti, migliorando l'efficienza e l'efficacia dell'intero percorso diagnostico e terapeutico. L'obiettivo del progetto è lo sviluppo di un sistema per la **segmentazione automatica** dei polipi presenti sulla parete del colon in immagini di colonscopia. Nei seguenti paragrafi si discuteranno le scelte che hanno portato alla strategia finale adottata, riportata in Fig 1.

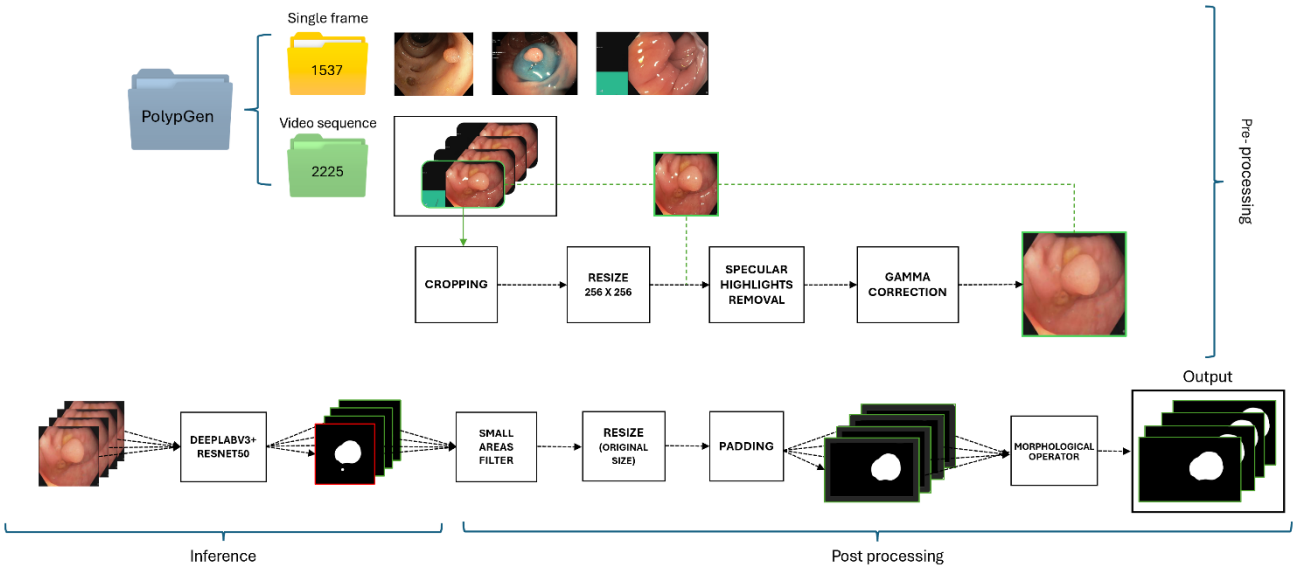

Fig 1. Rappresentazione grafica della pipeline finale proposta

2 Materiali e metodi

2.1 Dataset

I dati di partenza utilizzati per questo studio provengono da **PolypGen** [4], un dataset open-access contenente sia frame singoli che sequenze video di colonscopia, raccolti da sei diversi centri europei e africani. In particolare, il dataset fornito comprende: 1.537 **frame singoli**, tutti positivi (contenenti polipi), suddivisi in sei cartelle, una per ciascun centro di provenienza; e 2.225 frame estratti da 23 **sequenze** video positive, che includono sia frame positivi che frame negativi, distribuiti in 23 cartelle, ciascuna corrispondente a una sequenza. In Fig 2 sono mostrate alcune immagini del dataset. Ogni insieme di immagini è accompagnato da un corrispondente set di maschere di segmentazione, annotate da sei gastroenterologi esperti. Queste maschere forniscono una segmentazione dettagliata delle lesioni identificate nelle immagini, garantendo un'accurata delineazione delle aree di interesse.

2.2 Metriche per la valutazione dei risultati

Per valutare le prestazioni del modello nella segmentazione dei polipi, sono state utilizzate diverse metriche standard e specifiche per l'analisi degli errori. Le metriche principali scelte per la valutazione del modello includono:

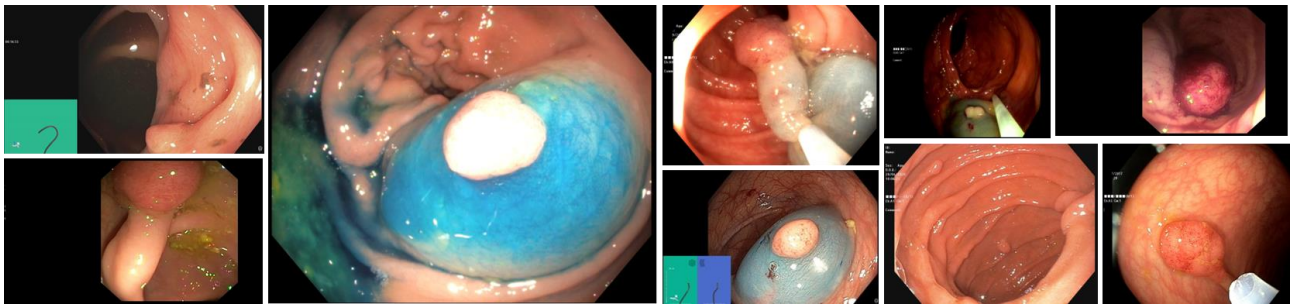

Fig 2. Immagini di esempio nel dataset

Dice Similarity Coefficient (DSC) e Intersection Over Union (IoU) sono due metriche ampiamente utilizzate per valutare la **qualità della segmentazione** misurando la sovrapposizione tra la segmentazione predetta e quella di riferimento. Il valore è compreso tra 0 e 1: un valore di 1 significa che l'area segmentata dal modello coincide perfettamente con l'area di riferimento (ground truth), un valore di 0 invece significa che non c'è alcuna sovrapposizione.

$$DSC = \frac{2\,|Area_{Pred} \cap Area_{GT}|}{|Area_{Pred}| + |Area_{GT}|} \qquad (1)$$

$$IoU = \frac{|Area_{Pred} \cap Area_{GT}|}{|Area_{Pred} \cup Area_{GT}|} \qquad (2)$$

DSC e IoU, tuttavia, non sono in grado di distinguere tra sovra-segmentazione e sotto-segmentazione, per cui è necessario combinarle con altre metriche, come Precision e Recall. La **precision** misura la proporzione di pixel correttamente identificati come polipo rispetto al totale delle predizioni positive (3). La **recall** misura la capacità del modello di individuare correttamente tutte le regioni contenenti polipi (4). Anche per precisione e recall il valore è compreso tra 0 e 1: una precision pari a 1 indica che tutti i pixel identificati come parte di polipi dal modello lo sono effettivamente, mentre se nessuno di essi è realmente parte di un polipo si ha una precision nulla. Un valore di 1 in Recall significa che il modello ha identificato correttamente tutti i pixel relativi ai polipi presenti nell'immagine, mentre 0 indica che il modello non ha rilevato alcun pixel del polipo, commettendo un errore totale.

$$Precision = \frac{TP}{TP+FP} \qquad (3)$$

$$Recall = \frac{TP}{TP+FN} \qquad (4)$$

Oltre alle metriche appena descritte, sono stati definiti indicatori specifici per analizzare in dettaglio gli **errori** del modello. Il **Relative Area Error (RAE)** misura la differenza relativa tra l'area segmentata dal modello e l'area reale del polipo, quantificando la **sovrastima** (valore positivo) o **sottostima** (valore negativo) delle dimensioni del polipo da parte del modello secondo la formula 5.

$$RAE = \frac{Area_{Pred} - Area_{GT}}{Area_{GT}} \qquad (5)$$

Il **Non-Matching Object Error (NMOE)**, è stato utilizzato per individuare i casi di **frame falsi positivi** nella segmentazione, ossia quando il modello segmenta oggetti che non corrispondono ai polipi reali.

2.3 Data Cleaning and Preparation

Inizialmente è stata condotta un'analisi della distribuzione delle classi, considerando congiuntamente sia i frame singoli sia quelli estratti dalle sequenze video. L'analisi ha evidenziato la presenza di 3122 frame positivi e 640 frame negativi, con un rapporto **positivi-negativi di 4.88**. A causa di questo forte sbilanciamento, si è deciso di **integrare ulteriori frame negativi** selezionati da sequenze negative del dataset PolypGen [4], al fine di migliorare la capacità del modello nel discriminare tra polipi reali ed elementi che potrebbero essere erroneamente classificati come tali, come le anse intestinali. Per mantenere un controllo sulla qualità dei dati, non sono state aggiunte intere sequenze negative, ma solo i frame più informativi, selezionati da sei specifiche sequenze (2, 8, 9, 18, 20 e 23) in base al livello di sfocatura calcolato utilizzando la varianza del Laplaciano. Seguendo lo stesso criterio, sono stati rimossi 211 frame positivi e 13 frame negativi a causa di una qualità insufficiente dell'immagine. Queste operazioni hanno portato a una nuova distribuzione del dataset, composta da **2911 frame positivi** e **1137 frame negativi**, con un rapporto **positivi-negativi ridotto a 2.56.**

2.4 Pre-processing

2.4.1 Cropping

Il processo di pre-processing inizia con l'implementazione di una funzione di cropping delle immagini per ridurre le variazioni e concentrare l'attenzione del modello sulle aree più significative, seguendo la pipeline riassunta in Fig 3. Questo passaggio si è reso necessario a causa della notevole variabilità riscontrata che potrebbe confondere il modello durante la fase di training, rendendo più difficile apprendere le caratteristiche rilevanti. I principali problemi riscontrati nelle diverse immagini sono stati:

- Rettangoli colorati (verde, grigio o blu) posizionati molto vicini o sopra l'oggetto d'interesse.
- Sfondi neri non completamente uniformi che presentano variazioni nei valori di pixel e forti rumori di fondo (visibili nello spazio HSV).
- Sottile linea bianca lungo il margine superiore o inferiore e testo bianco in prossimità della scena.
- Immagini con miniature o aree secondarie nello stesso frame, che possono confondere il modello.
- Regione di interesse molto scura, che risulta difficile da distinguere dal fondo.

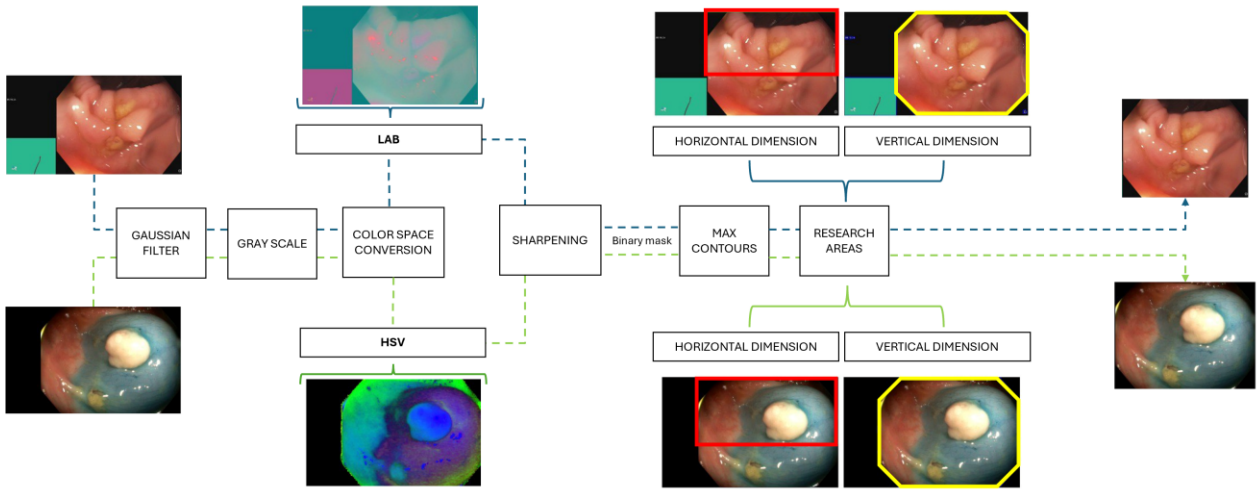

Fig 3. Pipeline delle fasi di cropping

Abbiamo inizialmente esplorato lo spazio colore RGB come riferimento per il cropping. Tuttavia, in questo spazio non è stato possibile definire soglie adeguate per un cropping efficace, a causa dell'elevata complessità cromatica delle immagini. Questa complessità deriva dalla presenza di colori simili tra gli elementi di interesse e lo sfondo: la mucosa appare rosa/rossa, i polipi possono essere bianchi, rosa, rossi, verdi o blu, mentre il canale del colon in profondità presenta aree nere, spesso indistinguibili dallo sfondo. L'utilizzo dello spazio colore **HSV** ha permesso di isolare in modo più efficace lo sfondo dall'area di interesse. Tuttavia, in alcuni casi specifici, il rumore presente in sfondi non omogenei risultava particolarmente marcato, rendendo difficile una separazione accurata. Per questo motivo, è stato adottato lo spazio colore **LAB**, che si è dimostrato meno sensibile a questo tipo di rumore. Tuttavia, lo spazio LAB non si è rivelato ottimale per immagini in cui la sezione di interesse era molto scura, come nella sequenza 5 positiva, dove il polipo è spesso posizionato ai margini della scena e in ombra. Per ottenere un cropping efficace, è stata quindi adottata una combinazione dei due spazi colore (LAB e HSV), utilizzando ciascuno in base al superamento di una soglia predefinita.

1. **Gaussian Filter e selezione dello spazio colore:**

   Come operazione preliminare è stato necessario ridurre il significativo rumore nello sfondo, utilizzando un **filtro gaussiano** con un kernel 7x7 e sigma pari a 5. Per selezionare lo spazio colore, l'immagine viene convertita in **scala di grigi** e calcolato il valore medio dell'intensità luminosa. Sulla base di questo valore, si confronta la luminosità dell'immagine con una soglia predefinita (threshold_space) fissata a 87. Se il valore medio è maggiore o uguale alla soglia, l'immagine viene convertita nello spazio **LAB**, altrimenti nello spazio **HSV**.

2. Applicazione dello sharpening

Dopo aver trasformato l'immagine RGB nello spazio colore selezionato (LAB o HSV), viene applicato un **filtro di sharpening** per aumentare il contrasto tra le diverse regioni dell'immagine (specialmente i contorni). Il kernel 3x3 presenta un valore centrale che varia al variare dello spazio colore ed è definito come segue:

$$\begin{bmatrix} -1 & -1 & -1 \\ -1 & val\_sharp & -1 \\ -1 & -1 & -1 \end{bmatrix}$$

3. Generazione della maschera binaria

Dopo l'applicazione del filtro di sharpening, si procede alla creazione di una maschera binaria che consente di separare le regioni di interesse dallo sfondo. Per ottenere questa maschera, si seleziona il canale dell'immagine in base allo spazio colore adottato: se l'immagine è in LAB, si utilizza il canale L, che rappresenta la luminosità, mentre nello spazio HSV, si considera il canale H, responsabile della tonalità. Su questo canale viene quindi applicata una soglia di binarizzazione con un valore impostato a **18**. I pixel con valori superiori vengono assegnati a 1 (bianco – area di interesse), mentre quelli con valori inferiori vengono impostati a 0 (nero – sfondo), ottenendo così una prima segmentazione dell'immagine.

4. Applicazione degli operatori morfologici

Per migliorare la qualità della maschera binaria e garantire una segmentazione più accurata, vengono applicati diversi operatori morfologici, tutti eseguiti utilizzando un kernel di dimensione 5×5. Ognuno con un ruolo specifico nel processo di elaborazione. L'**apertura** viene implementata per eliminare la linea bianca visibile nel margine di alcune immagini, oltre a piccoli artefatti residui dopo la binarizzazione, in particolare quelli localizzati intorno ai rettangoli colorati presenti in alcune immagini. In altre con scarsa illuminazione e zone interne molto scure, la segmentazione in HSV tende a produrre un'immagine con oggetti separati. Per questo motivo, viene applicata una **dilatazione** più ampia (10 iterazioni) per far sì che gli oggetti si tocchino, facilitando successivamente l'applicazione di un operatore morfologico di **riempimento** per migliorare la continuità delle regioni segmentate. Nel caso dello spazio LAB, dove la segmentazione è già più compatta, sono sufficienti solo 2 iterazioni di dilatazione per preservare le regioni di interesse senza compromettere la scena. L'operatore di **erosione** viene applicato principalmente sulle immagini nello spazio **HSV**, consente di affinare la regione di interesse dopo la forte dilatazione (necessaria per unire gli elementi interni). Questo passaggio riduce la dimensione complessiva della segmentazione, evitando un'espansione eccessiva delle aree segmentate e migliorandone la definizione, garantendo una maggiore precisione nel cropping.

5. Identificazione del contorno più grande nella parte superiore dell'immagine:

Per individuare la **scena di interesse**, vengono identificati i contorni degli oggetti presenti nell'immagine e selezionato quello con l'**area maggiore**. Tuttavia, la presenza di **rettangoli di etichettatura** nella parte inferiore potrebbe interferire con il rilevamento dell'oggetto principale. Per evitare questo problema, l'analisi viene suddivisa in due fasi:

I. Determinazione della **larghezza**: viene analizzata solo la porzione **superiore** dell'immagine, corrispondente ai 3/5 dell'altezza totale, per identificare il contorno più grande e racchiuderlo in un bounding box. Questo consente di escludere le etichette presenti nella parte inferiore.
II. Determinazione dell'**altezza**: il contorno più grande viene individuato su tutta l'immagine, garantendo che la regione di interesse comprenda l'intero oggetto segmentato.

Le coordinate finali del ritaglio vengono ottenute combinando i due bounding box: la larghezza è determinata dall'analisi della parte superiore, mentre l'altezza è definita dal contorno più grande individuato nell'intera immagine. Se l'area risultante è troppo piccola rispetto all'immagine originale (si riduce della metà), si mantiene l'immagine intera, per evitare perdite di informazione.

Il cropping isola la regione di interesse, eliminando elementi di disturbo presenti nello sfondo, preservando la struttura dell'oggetto ed evitando di distorcere l'immagine con il successivo resize (Fig 4). Questa operazione consente al modello di focalizzarsi sulle caratteristiche essenziali durante il training e facilita la successiva fase di rimozione degli specular highlight.

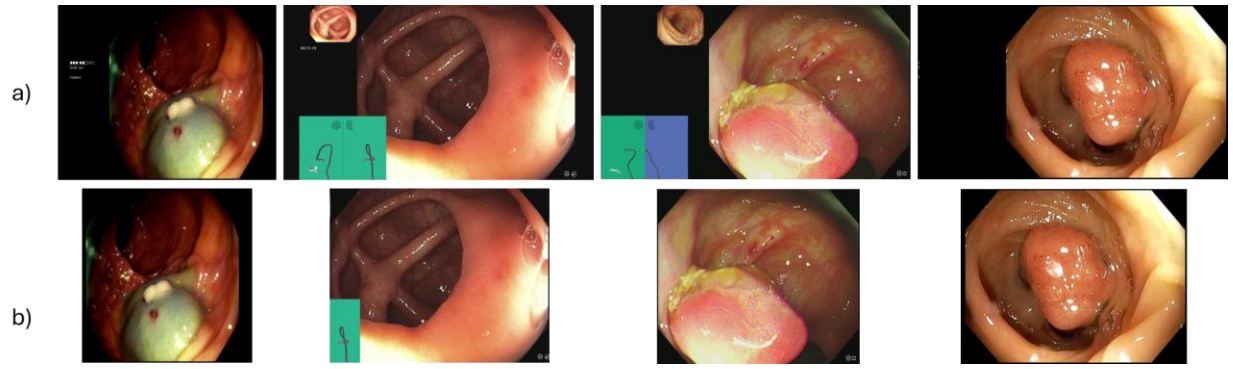

Fig 4. a) Immagini originali; b) Risultato del cropping

2.4.2 Resize

A causa della grande eterogeneità nelle dimensioni originali delle immagini (da minimo di 384x288 a un massimo di 1920x1080) e della necessità che l'input della rete abbia una risoluzione uniforme, è stato necessario applicare un ridimensionamento per uniformare tutte le immagini a una risoluzione di **256x256** pixel. Per il ridimensionamento delle **immagini**, è stata utilizzata l'**interpolazione bilineare**, che riduce il rumore e preserva la qualità visiva, mantenendo transizioni morbide tra i pixel. Il parametro anti_aliasing=True è stato impostato per minimizzare gli effetti di aliasing e migliorare ulteriormente la qualità. Per le **maschere** segmentate è stata prima eseguita una binarizzazione con una soglia di 128 per assicurare che i valori di pixel fossero limitati a 0 o 255, e successivamente è stata applicata l'**interpolazione nearest neighbor**, che preserva i confini delle regioni segmentate senza introdurre valori intermedi che potrebbero alterare le aree di interesse.

2.4.3 Specular highlights removal

Nelle immagini di colonscopia, la presenza di *specular highlights* rappresenta un problema significativo per l'analisi automatizzata, poiché queste regioni luminose possono compromettere l'accuratezza della segmentazione e della classificazione. Per affrontare questo problema, è stato sviluppato un metodo di pre-processing che identifica e rimuove i riflessi speculari attraverso una pipeline articolata in diversi passaggi, riportati in Fig 5.

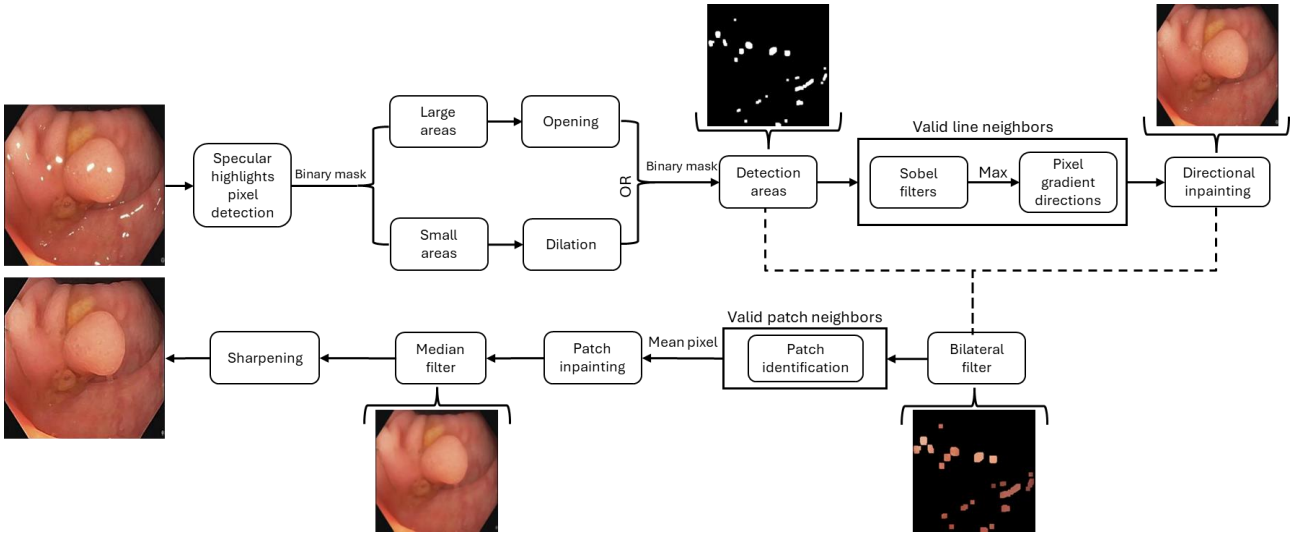

Fig 5. Passaggi effettuati per la rimozione di riflessi

1. **Specular highlights pixel detection**

   La prima fase consiste nell'identificare i pixel affetti da riflessi speculari. Per fare ciò, si analizzano i valori di intensità dei canali di colore, considerando che i riflessi appaiono tipicamente come aree con valori elevati nei canali rosso (R) e verde (G). Viene quindi creata una **maschera binaria**, dove un pixel viene classificato come riflesso se il valore di **R > 220** e **G > 190**. I pixel identificati vengono marcati con il valore 255 (bianco) nella maschera, mentre tutti gli altri sono impostati a 0 (nero).

2. **Filtraggio delle regioni in base alla dimensione**

Dopo aver ottenuto la maschera iniziale dei riflessi, è necessario distinguere tra aree di diversa estensione, poiché questo permette di applicare strategie di correzione specifiche e di evitare l'eliminazione di strutture rilevanti. In particolare, le regioni con un'estensione superiore a 1000 pixel vengono escluse dall'elaborazione, poiché alcuni frame potrebbero contenere polipi chiari di dimensioni simili. Le aree vengono quindi classificate come **piccole** se inferiori a **500 pixel** e **grandi** se comprese tra **500 e 1000 pixel**, consentendo un trattamento mirato in base alla loro dimensione.

3. **Refinement della maschera di riflessi**

Per migliorare l'affidabilità della maschera dei riflessi, si applicano operazioni morfologiche. La **dilatazione** (kernel 5x5) viene applicata alla maschera delle regioni più **piccole** per garantire che le aree rilevate siano più estese e comprendano eventuali bordi sfumati del riflesso. L'**apertura** (kernel 5x5) viene applicata alle aree più **grandi** per ridurre il rumore e migliorare la separazione tra le regioni di riflesso e le strutture circostanti. Il risultato di questa fase è una maschera raffinata che identifica con maggiore precisione le regioni affette da riflesso speculare, selezionando esclusivamente le aree su cui verrà applicato l'inpainting nelle fasi successive per la loro correzione.

4. **Compensazione delle regioni affette da riflesso: *Directional inpainting***

Dopo aver ottenuto una maschera raffinata, è necessario sostituire i pixel riflettenti con valori appropriati:

- Viene calcolata la **direzione del gradiente massimo** nell'immagine utilizzando operatori Sobel in quattro direzioni (0°, 45°, 90°, 135°), determinando così l'orientazione ottimale lungo la quale individuare pixel validi da utilizzare per la ricostruzione delle aree affette da riflesso.
- Per ogni pixel riflettente, si ricerca il primo **pixel non riflettente e non nero** (definito con una soglia > 50 su tutti e tre i canali RGB) lungo la direzione del gradiente.
- Il valore di questo pixel viene utilizzato per **sostituire il valore del pixel riflettente**.

Questo approccio permette di mantenere una transizione naturale nelle regioni ricostruite, evitando la creazione di artefatti evidenti.

5. **Filtraggio selettivo per la riduzione degli artefatti**

Per migliorare ulteriormente la qualità delle aree ricostruite, viene applicato un **filtro bilaterale** utilizzando un kernel 9x9 e impostando a 75 le soglie di similarità dei colori (sigmaColor) e di distanza spaziale (sigmaSpace). Questo filtro viene utilizzato esclusivamente sulle regioni identificate dalla maschera dei riflessi, con l'obiettivo di attenuare il rumore e ridurre gli artefatti residui generati dalla prima fase di inpainting (directional inpainting), garantendo al contempo la preservazione dei dettagli strutturali dell'immagine.

6. **Compensazione degli artefatti: *Patch inpainting***

Dopo aver filtrato gli artefatti, è necessario eseguire una seconda fase di inpainting per garantire una maggiore omogeneità nelle aree trattate, migliorando la continuità visiva e riducendo eventuali discontinuità residue. A questo scopo, viene applicato un approccio basato su **patch locali** per la ricostruzione delle regioni riflettenti. Per ogni pixel mascherato come riflesso:

- Si estrae un **intorno locale** di dimensione **15×15** pixel (patch_size) attorno al pixel riflettente.
- Vengono considerati **solo i pixel validi**, ossia quelli **non riflettenti** e con valori RGB superiori a **25** per evitare di utilizzare pixel scuri o artefatti.
- Il valore del pixel riflettente viene **sostituito dalla media** dei pixel validi all'interno del *patch*, garantendo una transizione graduale e naturale con le aree circostanti

7. **Smoothing e Sharpening**

Dopo il filtraggio selettivo, viene applicato un **filtro mediano** (kernel 5x5) **su tutta l'immagine** per ridurre ulteriormente il rumore residuo e uniformare la transizione tra le aree ricostruite e quelle originali. L'ultima fase della rimozione dei riflessi prevede l'applicazione di un **filtro di sharpening** per migliorare la nitidezza dell'immagine e

ripristinare i dettagli strutturali attenuati dai filtri di smoothing precedenti. Per questo scopo, viene utilizzato un **kernel 3×3** con la seguente matrice:

$$\begin{bmatrix} -0.5 & -0.5 & -0.5 \\ -0.5 & 5 & -0.5 \\ -0.5 & -0.5 & -0.5 \end{bmatrix}$$

Questo kernel è stato scelto in quanto permette di enfatizzare i contorni e i dettagli dell'immagine in modo **leggero** senza introdurre eccessivo rumore, mantenendo un buon equilibrio tra nitidezza e qualità visiva complessiva.

Attraverso questa pipeline, i riflessi speculari vengono individuati e rimossi, migliorando la qualità visiva delle immagini, visibile in Fig 6. Tuttavia, presenta alcune limitazioni: riflessi di grandi dimensioni non vengono compensati per evitare la rimozione errata di polipi chiari, mentre in immagini molto sfocate la rilevazione dei riflessi risulta meno accurata. Inoltre, la compensazione può essere meno uniforme nei riflessi vicini ai bordi, in aree scure o prossime alla strumentazione. La correzione nelle zone con variazioni cromatiche significative, come rossori, sanguinamenti o colorazioni di indaco risulta meno omogenea rispetto a regioni più uniformi dell'immagine.

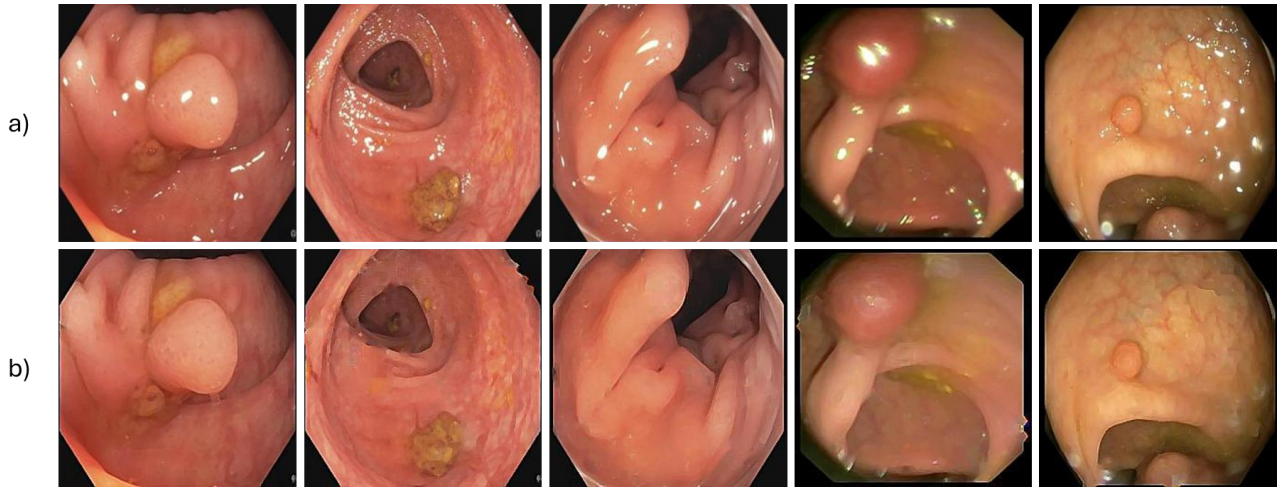

Fig 6. a) Immagini originali con riflessi; b) Risultato della rimozione di specular highlights

2.4.4 Gamma Correction

L'ultima fase del pre-processing prevede l'applicazione di una **correzione gamma** per migliorare la qualità visiva dell'immagine e bilanciare l'intensità luminosa. La correzione viene eseguita utilizzando un valore di gamma pari a **1.2**, che attenua leggermente il contrasto e **riduce l'impatto visivo delle aree molto luminose**, come i riflessi grandi non completamente eliminati durante il processo di rimozione di specular highlights. Le zone luminose perdono enfasi, rendendo l'immagine meno brillante, mentre le aree più scure non subiscono un'ulteriore penalizzazione, grazie al leggero sbilanciamento introdotto dalla correzione gamma. L'effetto complessivo dell'intera pipeline di preprocessing è mostrato in Fig 7.

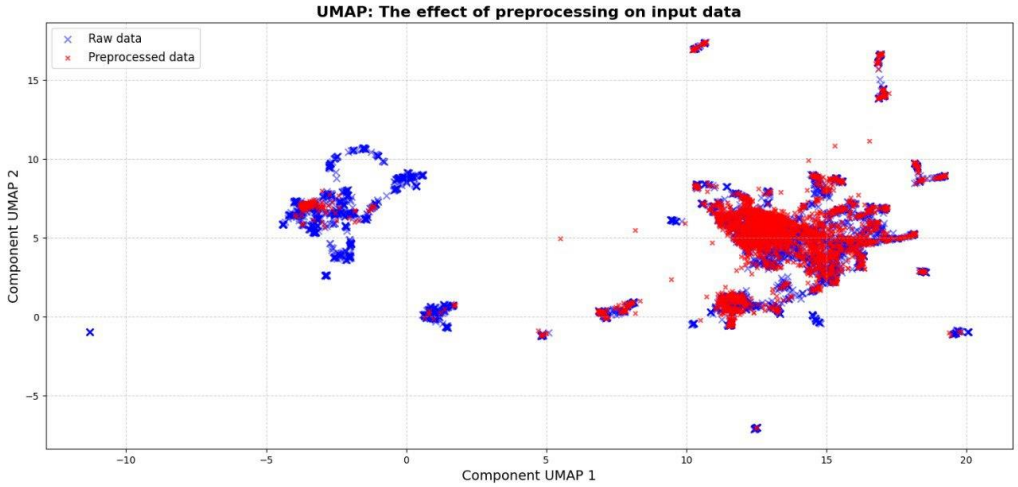

Fig 7. Effetto del preprocessing sui dati di input rappresentato tramite U-MAP

2.5 Divisione del Dataset

In un primo momento, il training e il validation set del modello sono stati costruiti esclusivamente con i frame singoli, poiché rappresentavano esempi più informativi e controllati per l'addestramento. Tuttavia, durante i test con le sequenze video, si è osservato che il modello faticava a segmentare correttamente molte di esse, a causa della maggiore variabilità presente in questi dati, come sfocature, variazioni di luminosità, oscurità e altri fattori. Per migliorare la capacità del modello di generalizzare su questi scenari più complessi, si è quindi deciso di ampliare il training e il validation set includendo alcune sequenze. In base a queste considerazioni, la suddivisione è stata eseguita separatamente per **frame singoli** e **sequenze video**, rispettando specifici criteri di distribuzione. I **frame singoli** sono stati suddivisi nelle proporzioni **80%** per il training set, **10%** per il validation set e **10%** per il test set. Questa divisione è stata effettuata garantendo un'equa distribuzione rispetto a:

- Risoluzione delle immagini
- Classe (positivi/negativi)
- Dimensione dei polipi (se presenti)
- Numerosità dei polipi nel frame

Per le **sequenze video** è stato necessario **mantenere tutti i frame appartenenti a una stessa sequenza all'interno dello stesso insieme** (training, validation o test), garantendo coerenza nei dati di addestramento. L'assegnazione è stata effettuata considerando i medesimi criteri utilizzati per i frame singoli, oltre a un'**analisi visiva** basata sulla presenza di mucosa, fluidi, secrezioni, vasi sanguigni, riflessi e strumenti e i livelli di luminosità e contrasto**.** In base a queste considerazioni, le **sequenze positive** sono state distribuite come segue: nel training set le sequenze 3, 5, 6, 7, 8, 13, 14, 15, 23; nel validation set le sequenze 1, 2, 16, 17, 20; e nel test le sequenze 4, 9, 10, 12, 19, 21, 22, 11, 18. Infine, per quanto riguarda le **sequenze negative**, essendo state aggiunte con l'obiettivo di migliorare la capacità discriminativa del modello, sono state inserite interamente nel **training set**.

2.6 Rete Neurale

La rete neurale utilizzata nella pipeline è **DeepLabV3+**, pre-addestrata sul dataset Cityscapes. Inizialmente, il modello è stato addestrato esclusivamente sui **singoli frame positivi** per adattarlo al contesto specifico del problema. Successivamente, la **baseline** è stata ottenuta addestrando DeepLabV3+ sull'**intero dataset**, composto da singoli frame positivi, sequenze video positive e sequenze video negative, fornendo così un riferimento per valutare le prestazioni delle configurazioni successive. Durante questa fase, sono state esplorate diverse configurazioni di **iperparametri** per ottimizzare le prestazioni:

- Backbone testati: ResNet50 e ResNet101
- Batch size: {6, 8}
- Loss function: Cross Entropy Loss con pesi {9:1, 8:2, 7:3}
- Ottimizzatori:
- AdamW con weight decay {0.01, 0.001, 0.0001} e learning rate {0.001, 0.0001, 5e-05}
- SGD con weight decay = 0.0005 e learning rate {0.01, 0.001, 0.0001}
- Numero di epoche: variabile, determinato dal criterio di early stopping, che interrompe l'addestramento se il mDice sul set di validazione non mostra miglioramenti dopo 20 iterazioni consecutive

Queste diverse configurazioni hanno permesso di valutare l'impatto delle diverse strategie di addestramento sul modello, ottimizzando la segmentazione sia dei frame che delle sequenze video. La configurazione finale selezionata prevede **ResNet50** come backbone, **batch size 6**, **Cross Entropy Loss** come funzione di perdita e **SGD** come ottimizzatore. Il numero di epoche è stato determinato in base alle prestazioni massime raggiunte sul set di validazione, risultando in **30 epoche**. Il learning rate iniziale è stato impostato a 0.01, per poi essere ridotto prima a 0.001 e successivamente a 0.0001, una volta raggiunto un checkpoint oltre il quale non si osservavano ulteriori miglioramenti nel set di validazione.

2.7 Post-processing

Il post-processing si è reso necessario per migliorare le predizioni del modello, in particolare per ridurre i **falsi positivi** e correggere la **sotto-segmentazione** osservata nella maggior parte delle sequenze. Analizzando visivamente le predizioni, è emerso che molti falsi positivi erano causati da segmentazioni errate dovute al rumore residuo presente nelle immagini, e la segmentazione risultava nettamente inferiore rispetto alle dimensioni dei polipi. Per affrontare questi problemi, il

post-processing è stato strutturato in tre fasi: 1) **riempimento** delle discontinuità all'interno delle segmentazioni per garantire una copertura più completa delle aree segmentate; 2) **eliminazione** delle regioni di dimensione inferiore a una soglia predefinita, al fine di mitigare i falsi positivi generati dal rumore; 3) **dilatazione** morfologica delle regioni segmentate per compensare la sotto-segmentazione e migliorare la coerenza spaziale della segmentazione. Questa sequenza di operazioni ha permesso di affinare le maschere predette, riducendo il rumore e migliorando l'aderenza delle segmentazioni alla ground truth.

3 Risultati Ottenuti

In questo paragrafo vengono riportati i risultai ottenuti con le varie implementazioni di pre-processing, tuning degli iperparametri e post-processing. Il pre-processing di solo resize è stato considerato come **baseline** della pipeline, servendo da punto di riferimento per valutare l'impatto delle successive trasformazioni. La Tabella 1 riassume i risultati ottenuti con diverse strategie di pre-processing. Si osserva che l'integrazione progressiva di operazioni come cropping, rimozione dei riflessi e gamma correction porta a miglioramenti significativi delle metriche di segmentazione.

Pre-processing	**IoU**	**DSC**	**Precision**	**Recall**	**NMOE**	**RAE**
res	42,52 ± 15,87	47,98 ± 15,71	66,42 ± 18,57	43,77 ± 16,31	**5,44**	-49,58 ± 20,99
crop_res	55,32 ± 25,91	61,22 ± 26,33	76,67 ± 25,09	56,72 ± 26,25	6,56	-36,42 ± 23,37
crop_res_specH	60,44 ± 22,85	68,24 ± 21,8	78,23 ± 37,21	64,65 ± 30,25	6,67	-20 ± 21,79
crop_res_specH_gamma	**62,2 ± 20,68**	**69,59 ± 19,31**	**79,84 ± 16,44**	**66,45 ± 19,99**	6,44	**-18,37 ± 20,41**

Tabella 1. Risultati dello studio di ablazione sulle diverse configurazioni di pre-processing

La Tabella 2 confronta le prestazioni di ResNet50 e ResNet101 come backbone della rete DeepLabV3+. L'utilizzo di **ResNet50** ha prodotto risultati migliori rispetto a ResNet101, ed è il backbone utilizzato per tutte le altre valutazioni.

Backbone	**IoU**	**DSC**	**Precision**	**Recall**	**NMOE**	**RAE**
ResNet50	**62,2 ± 20,68**	**69,59 ± 19,31**	**79,84** ± 16,44	**66,45 ± 19,99**	6,44	**-18,37 ± 20,41**
ResNet101	59,56 ± 21,85	65,32 ± 21,02	77,02 ± 15,03	61,84 ± 22,46	**0**	-28,04 ± 21,81

Tabella 2. Confronto performance al variare del Backbone

In Tabella 3, vengono riportati i risultati relativi alla modifica della dimensione del batch. L'aumento del batch da 6 a 8 ha comportato un peggioramento delle prestazioni; quindi, è stato mantenuto un **batch size** di **6**.

Batch size	**IoU**	**DSC**	**Precision**	**Recall**	**NMOE**	**RAE**
6	**62,2 ± 20,68**	**69,59 ± 19,31**	79,84 ± 16,44	**66,45 ± 19,99**	6,44	**-18,37** ± 20,41
8	59,57 ± 21,06	66,17 ± 20,5	**79,9 ± 18,44**	62,46 ± 21,06	**5,89**	-27,85 ± 20,88

Tabella 3. Confronto performance al variare del batch size

La Tabella 4 mostra il confronto tra diverse configurazioni di **pesi** della funzione di perdita Cross Entropy Loss. Nonostante un incremento di falsi positivi, la configurazione **90-10** ha portato a risultati migliori rispetto all'opzione 80-20.

Loss function weights	**IoU**	**DSC**	**Precision**	**Recall**	**NMOE**	**RAE**
90-10	**60,85 ±** 15,86	**67,29 ±** 15,16	**79,84 ±** 16,44	**66,45 ±** 19,99	10,33	**-21,32 ±** 20,11
80-20	60,2 ± 15,59	66,7 ± 15,11	72,71± 7,30	51,68± 16,35	**7,37**	-27,45 ± 17,88

Tabella 4. Confronto performance al variare dei pesi della funzione di loss

Per ridurre il numero di frame falsi positivi (NMOE pari a 6.44) riportati nella Tabella 3, è stato implementato un post-processing per migliorare la segmentazione. La Tabella 5 presenta i risultati ottenuti applicando una prima fase di post-elaborazione, che comprende l'operatore morfologico **fill holes** per colmare le discontinuità nelle segmentazioni e la **rimozione delle aree di piccole dimensioni** per ridurre i falsi positivi. Inoltre, è stata applicata un'operazione morfologica di **dilatazione** con un elemento strutturale a forma di disco di raggio 3, testando da 1 a 3 iterazioni. Il miglior risultato è stato ottenuto con **due iterazioni**. La Tabella 6 riporta i risultati del post-processing sul set di validazione.

Nel complesso, l'ottimizzazione del post-processing ha permesso di migliorare significativamente la segmentazione, bilanciando la riduzione dei falsi positivi e la correzione della sotto-segmentazione, contribuendo a un incremento delle prestazioni globali del modello.

TEST SET						
Post-processing	**IoU**	**DSC**	**Precision**	**Recall**	**NMOE**	**RAE**
no	62,91 ± 20,91	70,1 ± 19,45	79,28 ± 16,4	67,72 ± 20,54	6,56	-15,8 ± 21,24
Fill + filter small area	63,89 ± 20,3	71,17 ± 18,84	**80,37 ± 16,49**	68,12 ± 19,35	6,00	-18,64 ± 20,36
Fill + filter small area + dilatazione (disk=3, iter=2)	**64,44 ± 19,78**	**71,51 ± 17,98**	76,32 ± 16,31	**71,9 ± 19,52**	**3,22**	**-9,34 ± 22,97**

Tabella 5. Risultati del post-processing sul test set

VALIDATION SET						
Post-processing	**IoU**	**DSC**	**Precision**	**Recall**	**NMOE**	**RAE**
no	58,62 ± 24,78	62,82 ± 27,57	67,14 ± 29,58	62,65 ± 28,23	8,80	-10,05 ± 21,48
Fill + filter small area + dilatazione (disk=3, iter=2)	**62,37 ± 23,18**	**66,45 ± 25,83**	**68,17 ± 26,5**	**67,87 ± 27,53**	**6,80**	**-5,74 ± 22,92**

Tabella 6. Risultati del post-processing sul validation set

4 Limitazioni

4.1 Problemi riscontrati dovuti al dataset

Durante lo sviluppo della pipeline, sono emerse diverse difficoltà legate alle caratteristiche del dataset. Nella fase di cropping la grande variabilità tra le immagini ha rappresentato una sfida significativa. Differenze nelle dimensioni complessive, nella grandezza e posizione del polipo, nonché la presenza di aree molto scure e rumore di fondo, hanno reso complesso definire una strategia di ritaglio uniforme ed efficace. Nel processo di rimozione degli specular highlights la presenza di riflessi molto estesi e intensi, che coprono parti rilevanti dell'oggetto, ha reso difficile la loro eliminazione senza compromettere i dettagli essenziali per la segmentazione. In molti casi, la rimozione completa di questi artefatti ha comportato una perdita di informazioni visive importanti. Inoltre, sono stati individuati errori e anomalie nelle maschere manuali fornite, per esempio: 1) segmentazioni estese oltre la scena, includendo erroneamente lo sfondo; 2) frame consecutivi con la stessa visualizzazione ma segmentazioni manuali incoerenti, introducendo incertezza nei dati di addestramento e nella valutazione del modello. Queste problematiche hanno influenzato la qualità del pre-processing e della segmentazione, rendendo necessaria l'adozione di strategie adattative per migliorare la coerenza dei risultati.

4.2 Criticità della soluzione

Il processo di cropping non ha permesso di ottenere un ritaglio uniforme tra tutte le immagini, principalmente a causa delle problematiche discusse in precedenza, come la variabilità nelle dimensioni, la posizione del polipo e la presenza di aree scure o rumorose. Inoltre, la scelta di calcolare la dimensione orizzontale del ritaglio basandosi su una sotto-area dell'immagine ha limitato l'applicabilità della funzione al nostro dataset, riducendo la generalizzabilità del metodo in altri contesti. Per quanto riguarda la rimozione degli specular highlights, la soglia impostata per evitare di eliminare riflessi di grandi dimensioni, e quindi non interferire con la segmentazione del polipo, ha impedito una rimozione completa di questi artefatti, lasciando residui che in alcuni casi coprono parti rilevanti dell'immagine. Inoltre, il processo di inpainting non è riuscito a garantire una correzione uniforme: sebbene il rumore residuo venga attenuato, in alcune aree rimane comunque visibile. Infine, la complessità cromatica di diversi frame ha reso difficile applicare un inpainting omogeneo e fedele, introducendo variazioni che possono compromettere la qualità della segmentazione finale.

5 Sviluppi futuri

Un possibile miglioramento consiste nell'ottimizzazione del cropping per rimuovere completamente i bordi residui, evitando imprecisioni nell'inpainting dei riflessi dovute a una rimozione incompleta. Inoltre, data la complessità e variabilità del riflesso, si potrebbe integrare l'uso di algoritmi di apprendimento automatico, come reti neurali specializzate (U-Net o GAN), per una rimozione più accurata rispetto ai metodi basati su soglie. Infine, considerando la natura video del problema, sarebbe utile sperimentare architetture come le Recurrent Neural Networks per sfruttare la correlazione temporale e ottenere segmentazioni più stabili nel tempo.

6 Bibliografia

[1] E. Morgan *et al.*, 'Global burden of colorectal cancer in 2020 and 2040: incidence and mortality estimates from GLOBOCAN', *Gut*, vol. 72, no. 2, pp. 338–344, Feb. 2023, doi: 10.1136/gutjnl-2022-327736.

[2] M. Bretthauer *et al.*, 'Effect of Colonoscopy Screening on Risks of Colorectal Cancer and Related Death', *N. Engl. J. Med.*, vol. 387, no. 17, pp. 1547–1556, Oct. 2022, doi: 10.1056/NEJMoa2208375.

[3] E. Young, L. Edwards, and R. Singh, 'The Role of Artificial Intelligence in Colorectal Cancer Screening: Lesion Detection and Lesion Characterization', *Cancers*, vol. 15, no. 21, p. 5126, Oct. 2023, doi: 10.3390/cancers15215126.

[4] S. Ali *et al.*, 'A multi-centre polyp detection and segmentation dataset for generalisability assessment', *Sci. Data*, vol. 10, no. 1, p. 75, Feb. 2023, doi: 10.1038/s41597-023-01981-y.