A Two-Stage Framework for Abnormalities Detection in WCE Images by Combining Semantic Segmentation and Deformable Agent-Based Classification

Brahim Alibouch, Yasmina El Khalfaoui
LabSIV-Department of Computer Science-Faculty of Sciences, Ibn Zohr University, Agadir, Morocco

***Abstract*—Wireless capsule endoscopy (WCE) has revolutionized gastrointestinal (GI) diagnostics by offering a patient-friendly imaging and diagnostic tool compared to traditional endoscopic techniques. However, the manual assessment of these images is a time-consuming task and is prone to inaccuracies, which necessitates the implementation of automated approaches. In this paper, we introduce a two-stage deep learning framework to identify the most common GI abnormalities in WCE images. The first stage in the proposed method is to segment suspicious regions from the WCE images, which act as potential markers for GI abnormalities. In the second stage we perform frame-level classification to identify and categorize different pathologies in the GI tract. Extensive experiments conducted on four image datasets demonstrate that our approach achieves the highest values in terms of accuracy, precision, recall and specificity in comparison with four common deep learning methods : Resnet50, VGG16, Vit-S16 and InceptionV3.**

Keywords*—Wireless capsule endoscopy; deep learning; classification; gastrointestinal abnormalities

I. Introduction

Wireless Capsule Endoscopy has emerged as a revolutionary non-invasive diagnostic technology for visualizing the GI tract, particularly the small bowel, which remains challenging to access through conventional endoscopic procedures. This innovative technique involves patients swallowing a pill-sized capsule equipped with cameras that capture thousands of images as it traverses the digestive system, providing comprehensive visualization without the need for sedation or exposure to ionizing radiation [1]. The clinical significance of WCE is particularly pronounced in the diagnosis of small bowel disorders, including suspected bleeding, iron deficiency anemia, and Crohn's disease, where it serves as a first-line diagnostic tool.

However, the clinical adoption of WCE presents significant challenges that limit its full potential in routine gastroenterological practice. Each WCE examination generates an overwhelming volume of data, typically producing 50,000 to 80,000 images per patient, with only 5-10% containing pathologically relevant findings [2]. This massive data burden places substantial strain on healthcare professionals who must manually review and analyze each frame, a process that is not only time-consuming but also prone to human error and fatigue. The situation is further complicated by the fact that 20-30% of captured frames may be obstructed by bubbles, food residues, or other contaminants, making accurate diagnosis even more challenging [3].

Among the various pathological findings detectable through WCE, polyps represent a particularly important category of lesions that require precise identification and characterization. Polyps, which are abnormal tissue growths protruding from the intestinal wall, can range from benign lesions to pre-cancerous or malignant formations. Their early detection and accurate classification are crucial for determining appropriate treatment strategies and preventing progression to more serious conditions. However, polyp detection in WCE images presents unique technical challenges due to their variable size, morphology, and appearance, as well as the inherent limitations of capsule imaging, including motion artifacts, varying illumination conditions, and the presence of intestinal contents that may obscure visualization.

The integration of artificial intelligence (AI) and deep learning technologies in WCE analysis has shown tremendous promise in addressing these challenges. Recent advances in computer vision and machine learning have demonstrated the potential to reduce reading time by up to 95% while improving diagnostic accuracy [1]. Deep learning models, particularly convolutional neural networks (CNNs), have proven superior to classical machine learning methods due to their end-to-end learning capabilities and automatic feature extraction mechanisms [4]. These technologies offer the potential to not only automate the detection process but also provide objective, reproducible results that can support clinical decision making.

Despite significant progress in applying deep learning to WCE image analysis, most existing approaches focus on either classification or segmentation tasks independently. Classification methods excel at identifying the presence of pathological conditions but provide limited spatial information about lesion location and boundaries. Conversely, segmentation approaches can precisely delineate lesion boundaries but may not provide comprehensive diagnostic information about lesion characteristics and severity. This separation of tasks represents a significant limitation in clinical practice, where both spatial localization and diagnostic classification are essential for comprehensive patient care.

The development of unified framework that simultaneously perform both segmentation and classification represents a critical advancement in WCE image analysis. Such integrated approaches can leverage the complementary strengths of both

tasks, where segmentation provides precise spatial localization that can inform and improve classification accuracy, while classification results can guide and refine segmentation boundaries. This synergistic relationship is particularly important for polyp analysis, where accurate boundary delineation is essential for size estimation and morphological assessment, while classification provides crucial information about polyp type and potential malignancy risk. Furthermore, the clinical workflow in gastroenterology naturally requires both spatial and categorical information about detected lesions. Clinicians need to know not only whether a polyp is present (classification) but also its exact location, size, and morphological characteristics (segmentation). An integrated framework that provides both types of information simultaneously would significantly enhance clinical utility and diagnostic confidence while streamlining the review process.

This work presents a novel deep learning framework designed to address the dual challenges of GI abnormalities segmentation and classification in WCE images through an integrated approach that leverages the synergistic relationship between these complementary tasks. Our framework aims to provide comprehensive diagnostic support that meets the practical needs of clinical gastroenterology while addressing the technical challenges inherent in WCE image analysis.

This paper is organized as follows. In the next section the related works are reviewed. In Section III, we explain the proposed method. Section IV presents the outcomes of the experiments conducted to evaluate our approach. Finally, Section V delineates the conclusions of this paper.

II. Related Works

The application of deep learning techniques to wireless capsule endoscopy has experienced rapid growth over the past decade, with researchers exploring various approaches to automate the detection, classification, and segmentation of GI abnormalities. The foundation of automated WCE analysis lies in the development of robust classification systems capable of distinguishing between normal and pathological findings. Soffer et al. [4] conducted a systematic review and meta-analysis of deep learning applications in WCE, revealing exceptional performance across various pathological conditions. Their analysis demonstrated that CNNs achieved high pooled sensitivity and specificity scores, with particularly impressive results for bleeding detection (sensitivity 0.98, specificity 0.99) and ulcer identification (sensitivity 0.95, specificity 0.94). The superior performance of CNNs over classical machine learning methods was attributed to their end-to-end learning capabilities and automatic feature extraction mechanisms. Recent advances in CNN architectures have shown remarkable improvements in classification accuracy. Habe et al. [5] provided a comprehensive review of deep learning performance across 30 studies from 2013-2022, revealing that Xception models achieved the highest accuracy (99%) on datasets containing 53,555 images, followed by DenseNet121 with 98.17% accuracy on only 1,000 images, demonstrating the effectiveness of modern architectures in achieving high performance even with limited training data. Transfer learning approaches, particularly those utilizing ImageNet pre-trained models, have consistently shown significant performance improvements across different architectures, with Inception-ResNet-v2 achieving 98.1% accuracy on 400,000 images.

The integration of attention mechanisms has further enhanced classification performance. Bandopadhyay et al. [6] demonstrated the superiority of Vision Transformer (ViT) architectures over traditional CNNs for GI ulcer classification in WCE images. Their ViT model achieved perfect scores across all key metrics (Accuracy, Precision, Recall, and F1-Score) when compared against prominent deep learning models including ResNet50, VGG19, Inception V3, and Inception-ResNet152V2, highlighting the potential of transformer-based architectures in medical image analysis. Several researchers have explored sophisticated approaches to address specific challenges in WCE classification. Garbaz et al. [7] proposed a dual-branch deep convolutional neural network combining DenseNet121 with residual blocks specifically designed for detecting small, indistinct vascular lesions such as angiodysplasia and bleeding zones. Their architecture achieved accuracies of 99.50% and 96.96% on MICCAI 2017 and Kvasir-Capsule datasets respectively, with the residual block component contributing 12.0% and 8.78% accuracy improvements. Li et al. [8] introduced a dynamic multi-task learning framework that addresses fine-grained recognition and class imbalance challenges through the combination of semi-hard triplet loss with weighted cross-entropy loss. Their approach achieved 96.47% F1-score on Kvasir-Capsule (10-class classification) and 96.75% F1-score on CAD-CAP (3-class classification), demonstrating the effectiveness of multi-objective optimization in handling complex classification scenarios. The challenge of limited labeled data has been addressed through various innovative approaches. Muruganantham and Balakrishnan [9] proposed ACT-WISE, an uncertainty-driven active learning framework integrated with deep semi-supervised learning. Their teacher-student consistency regularization methodology achieved 97% classification accuracy and 0.95 AUC on the Kvasir-Capsule dataset, outperforming traditional semi-supervised methods like Mean Teacher (95% accuracy) while significantly reducing annotation requirements.

Recognizing the inherent temporal nature of WCE videos, several researchers have explored spatiotemporal analysis approaches. Bordbar et al. [10] pioneered the application of three-dimensional CNNs for multiclass WCE classification, processing 16 consecutive frames to leverage temporal relationships that gastroenterologists naturally consider during diagnosis. Their 3D-CNN approach achieved significant improvements over 2D methods, with 98.92% sensitivity vs. 98.05%, 99.50% specificity vs. 86.94%, and 99.20% accuracy vs. 92.60%, demonstrating the value of incorporating temporal dependencies. Prakash and Krishnamurthy [11] developed an innovative transfer learning approach combining depth estimation with evolutionary optimization for optimal keyframe selection from WCE videos. Their TUMNetv2 architecture integrated transformer encoders with UNet+ and MobileNetv2, achieving 93% accuracy, 90% precision, 85% recall, and 87% F1-score while significantly reducing computational burden through intelligent frame selection. The combination of deep features with traditional hand-crafted features has shown promising results in WCE analysis. Amiri et al. [12] proposed a hybrid approach combining deep features from pre-trained ResNet50 with hand-crafted color, shape, and texture features. Their method utilized the Expectation Maximization algorithm for

region of interest extraction and achieved 97.82% accuracy in detecting various abnormalities including ulcers, bleeding, angiodysplasia, lymphoid hyperplasia, and polyps. Caroppo et al. [13] developed a transfer learning-based expert system that combined features from ResNet50 and Inception V4 architectures. Their hybrid approach, utilizing minimum Redundancy Maximum Relevance feature selection and SVM classification, achieved 98.09% accuracy on KID Dataset 2 and 94.48% accuracy on MICCAI 2017 dataset, with strong cross-dataset generalization capabilities (91.46% and 93.28% accuracy).

While most WCE research has focused on classification tasks, segmentation approaches have gained increasing attention for their ability to provide precise spatial localization of abnormalities. He et al. [14] proposed MVANet, a Multi-view Attention Network designed to address critical challenges in endoscopic image segmentation, including boundary blurring, low contrast, and small lesion detection. Their architecture achieved state-of-the-art performance with mDice values of 93.1% and 94.9%, and mIoU values of 88.4% and 90.7% on Kvasir-SEG and CVC-ClinicDB datasets respectively. Charfi et al. [15] presented a novel approach using Dense-UNet for segmentation followed by a modified residual attention network for classification. Their method focused on detecting and segmenting red lesions, ulcers, and polyps, addressing the challenge of processing thousands of low-quality images in WCE videos. The quality of WCE images presents significant challenges for automated analysis systems. Sadeghi et al. [3] addressed this issue by creating the first publicly available multicentre dataset for clear and contaminated region segmentation in WCE images. Their dataset included pixel-level annotations for 17,593 images, categorizing contamination levels from 1-10 based on percentage of contaminated area, with contamination ranging from minimal (0.28% ± 1.29%) to nearly complete coverage (95% ± 0.0%), highlighting the critical need for robust preprocessing and quality assessment mechanisms in WCE analysis systems. The detection of rare or previously unseen pathologies represents a significant challenge in WCE analysis. Quindós et al. [16] addressed this through a novel patch-based self-supervised approach for out-of-distribution detection. Their three-stage methodology achieved AUROC scores ¿0.6 for detecting unseen pathologies including lymphangiectasia (0.767), foreign bodies (0.729), ulcers (0.795), and blood (0.677), demonstrating the potential of unsupervised approaches for identifying novel pathological conditions. Accurate localization within the GI tract is crucial for clinical interpretation of findings. Fan et al. [17] developed an automatic locating system for identifying the pylorus and ileocecal valve using ResNet-50-based classification combined with a novel sliding pin algorithm. Their approach achieved mean locating errors of 26 frames for the pylorus and 65 frames for the ileocecal valve, representing significant improvements over previous state-of-the-art methods. Object detection frameworks have been adapted for WCE analysis to provide both classification and localization capabilities. Xiao and Feng [2] developed an improved YOLOv3-based system for small intestinal lesion detection, introducing a novel labeling strategy that expanded lesion annotation areas to 1.33 times the original size. Their approach achieved 93.5% mAP with 4% improvement over standard YOLOv3, demonstrating better speed-accuracy balance while successfully identifying lesions from small to large sizes. Fernandes et al. [18] developed a deep learning-based methodology to automatically identify physician-annotated frames within WCE videos, with MobileNet achieving superior performance with 94% accuracy for first match identification and 98% accuracy for top-20 matches.

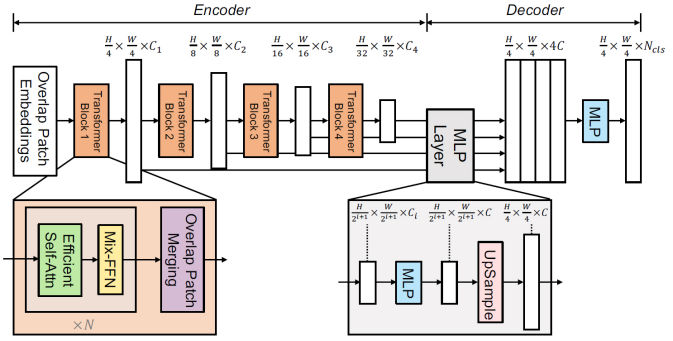

Fig. 1. SegFormer [19] architecture.

Despite significant progress in WCE image analysis, several limitations persist in current research. Most existing approaches focus on either classification or segmentation tasks independently, limiting their clinical utility where both spatial and categorical information are required. Additionally, while many studies report high accuracy on specific datasets, there remains a lack of comprehensive frameworks that address multiple aspects of polyp analysis simultaneously. The challenge of class imbalance, particularly evident in polyp detection where pathological cases are relatively rare, requires more sophisticated approaches that can effectively learn from limited positive examples while maintaining high sensitivity. Furthermore, the integration of clinical workflow considerations and real-time processing requirements has received limited attention in current research. The temporal nature of WCE videos presents both opportunities and challenges that have been insufficiently explored, and there remains significant potential for developing more sophisticated approaches that leverage the sequential nature of WCE data for improved diagnostic accuracy. Current research also lacks comprehensive evaluation frameworks that consider both technical performance metrics and clinical utility measures, making the development of standardized evaluation protocols that reflect real-world clinical scenarios essential for advancing the field toward practical clinical deployment. This comprehensive review highlights the significant progress made in applying deep learning to WCE analysis while identifying key gaps and opportunities for advancement. The proposed framework aims to address these limitations through an integrated approach that combines the strengths of both segmentation and classification methodologies for comprehensive polyp analysis in WCE images.

III. Methodology

In this paper, we introduce a two-stage deep learning framework to identify the most common GI abnormalities in WCE images. The first stage in the proposed method is to segment suspicious regions from the WCE images, which act as potential markers for GI abnormalities. To this end, we employed SegFormer-B0, a lightweight transformer-based semantic segmentation model. In the second stage we

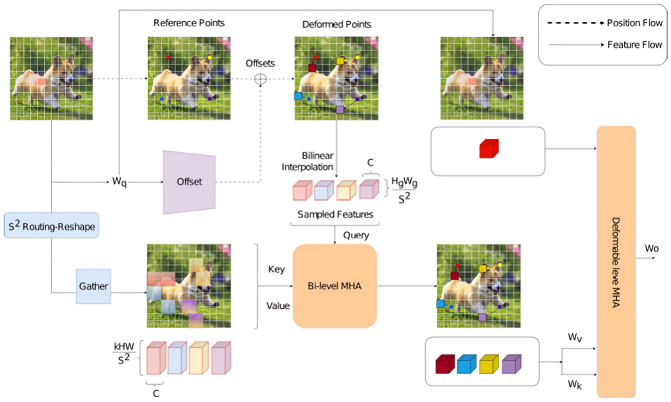

Fig. 2. Architecture of DBRA [20].

perform frame-level classification to identify and categorize different pathologies in the GI tract. For this purpose, we used DeBiFormer, a vision transformer architecture that uses an advanced attention mechanism called Deformable Agent Bi-level Routing Attention, which merges deformable attention with bi-level routing to reduce computational cost while improving accuracy.

The methodology combines the computational efficiency of SegFormer for initial segmentation with the advanced attention mechanisms of DeBiFormer for refined classification and abnormality detection.

A. SegFormer

Introduced by Xie et al. [19], SegFormer represents a significant step forward in applying Vision Transformers (ViTs) to semantic segmentation. The design philosophy of SegFormer focuses on creating a framework that is simple, efficient, and powerful, effectively overcoming the inherent limitations of prior ViT-based models in dense prediction tasks. It specifically addresses two major challenges: the low-resolution, fixed-size feature maps generated by standard ViTs, which are poorly suited for pixel-level classification, and the high computational complexity associated with self-attention, especially when dealing with high-resolution images.

The architecture consists of two main components: a hierarchical Mix Transformer (MiT) encoder and a lightweight All-MLP decoder as shown in Fig. 1. This design eliminates the need for positional encodings, supports multi-scale feature extraction, and significantly reduces model complexity while achieving state-of-the-art performance.

The encoder generates multi-scale features at resolutions of $\{1/4, 1/8, 1/16, 1/32\}$ of the original image, addressing the limitation of Vision Transformers that produce only single-scale representation and enabling the capture of both local and global features effectively, which is essential to handle dense prediction tasks like segmentation. The decoder takes the multi-scale feature maps from all four encoder stages and uses MLPs to transform and concatenate them together for pixel-wise predictions . The final output is a segmentation mask at $\frac{H}{4} \times \frac{W}{4} \times N_{cls}$ resolution, where Ncls represents the number of categories.

B. DeBiFormer

DeBiFormer (Deformable Bi-level Routing Attention Transformer) represents a significant advancement in vision transformer architectures by addressing key limitations in existing attention mechanisms. Introduced by Long et al. [20], DeBiFormer combines the flexibility of deformable attention with the semantic relevance of bi-level routing (Fig. 2) to create a more effective and interpretable attention mechanism for visual recognition tasks.

As shown in Fig. 3(a), the architecture of DeBiFormer follows a four-stage pyramid structure where each stage begins with either overlapped patch embedding (first stage) or patch merging modules (subsequent stages) to progressively reduce spatial resolution while increasing channel dimensions, followed by multiple consecutive DeBiFormer blocks. Each DeBiFormer block incorporates a 3×3 depthwise convolution for implicit relative position encoding, the DBRA module for cross-location relation modeling, and a 2-ConvFFN module for per-location feature embedding.

The core contribution of DeBiFormer is its novel Deformable Bi-level Routing Attention (DBRA) module [Fig. 3 (b)], which operates through an attention-in-attention architecture. Unlike traditional attention mechanisms that either focus on fixed patterns or lack semantic awareness, DBRA achieves both spatial flexibility and semantic relevance through a two-stage process:

- Deformable Point Generation: The module first generates deformable points using an offset network that shifts reference points toward informative regions based on query features. This allows the model to

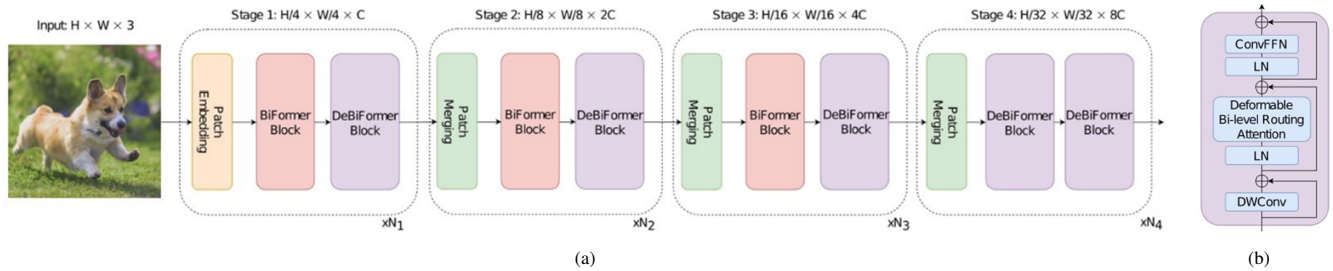

Fig. 3. (a) Overall DeBiFormer architecture, (b) DeBiFormer block [20].

adaptively focus on important areas rather than being constrained to regular grid patterns.

- Semantic Region Selection: Instead of attending to all deformed points equally, DBRA employs a bi-level routing mechanism where each deformable point acts as an agent query. These agents select only the top-k most semantically relevant regions, ensuring computational efficiency while maintaining semantic focus.

C. Proposed Method

This paper proposes a novel two-stage deep learning framework for the automated detection of abnormalities in WCE images. The overarching goal is to first precisely localize and segment potential pathological regions and subsequently classify them with high accuracy to reduce false positives. The proposed architecture, leverages the strengths of a state-of-the-art semantic segmentation model followed by a powerful, computationally efficient transformer-based classifier.

The first stage of our proposed framework is dedicated to the accurate segmentation of potential abnormalities within WCE images. For this task, we have selected SegFormer-B0, a lightweight vision transformer-based segmentation model that can effectively handle the hierarchical feature extraction required for medical image segmentation without the over-fitting issues common in traditional CNNs on small medical datasets. It processes the full WCE image and outputs a binary segmentation mask that precisely delineates the boundaries of suspicious regions. This stage acts as an intelligent attention mechanism, effectively filtering out vast areas of normal mucosa and focusing subsequent computational resources only on regions of clinical interest.

Following the segmentation of potential GI abnormalities by SegFormer-B0, the second stage of our framework involves classifying these segmented regions into specific pathological categories. For this classification task, we propose the use of DeBiFormer, leveraging its deformable agent bi-level routing attention mechanism to process both the original WCE images and the segmentation outputs from Stage 1. This dual-input approach enables the network to validate and refine segmentation results while providing robust abnormality classification. An overview of the proposed two-stage deep learning framework is illustrated in Fig. 4.

IV. Experiments

A. Datasets

For this work we used two datasets to pre-train the SegFormer-B0 model : CVC-ClinicDB [21] and Kvasir-SEG [22]. CVC-ClinicDB is a dataset consisting of 612 images extracted from colonoscopy video sequences, each with a corresponding binary ground truth mask highlighting polyps. All images are standardized at a resolution of 384 × 288 pixels, making the dataset compact and consistent. Kvasir-SEG, on the other hand, contains 1,000 images with varying resolutions, reaching up to 1920 × 1072 pixels, and each is paired with a high-quality expert-verified mask.

Once the SegFormer-B0 model pre-trained, two other datasets were used to train and test the whole proposed model. The first is Curated Colon Disease Dataset for Deep Learning [23] which includes 6000 WCE images of the GI tract, distributed across four categories: normal, ulcerative colitis, polyps, and esophagitis, with 1500 images per class. The second one is Kvasir v2 [24], a multi-class GI endoscopy dataset containing 8,000 endoscopic images evenly distributed across eight classes. Each class has 1,000 images, and the resolutions vary from 720 × 576 to 1920 × 1072 pixels.

Tables I and II provide detailed informations on each dataset.

TABLE I. Curated Colon Disease and Kvasir v2 Datasets Specifications

Dataset	Class	Train	Validation	Test	Total
Curated Colon Disease	Normal	1200	150	150	1500
	Ulcer	1200	150	150	1500
	Polyps	1200	150	150	1500
	Esophagitis	1200	150	150	1500
Kvasir v2	dyed-lifted-polyps	800	100	100	1000
	dyed-resection-margins	800	100	100	1000
	Esophagitis	800	100	100	1000
	normal-cecum	800	100	100	1000
	normal-pylorus	800	100	100	1000
	normal-z-line	800	100	100	1000
	polyps	800	100	100	1000
	ulcerative-colotis	800	100	100	1000

TABLE II. Kvasir-SEG and CVC-ClinicDB Datasets Specifications

Dataset	Train	Validation	Test	Total
Kvasir-SEG	800	100	100	1000
CVC-ClinicDB	490	61	61	612

B. Preprocessing and Data Augmentation

A combination of data augmentation and normalization techniques can improve the quality, diversity, and generalizability of training data [25]. In this work, the images are first resized to fixed size and several transformations are applied including rotations, vertical and horizontal flips, and affine transformations. Color jittering is performed to alter saturation contrast, hue and brightness also as random perspective distortion to enhance furthermore diversity. To maintain consistency during evaluation, only resizing and normalization are performed on validation set.

This preprocessing pipeline guarantees that the model encounters diverse and realistic image variations during training, while simultaneously maintaining the validation data in a standardized format for precise performance assessment.

C. Implementation Details

The SegFormer-B0 model was pre-trained with CVC-ClinicDB and Kvasir-SEG datasets using Adam optimizer and binary cross entropy loss function with a batch size of 4 over 60 epochs. The initial learning rate was set to 0.00005 and the process were executed on Google Colab, with an NVIDIA Tesla T4 GPU with 2560 CUDA cores and 16 GB of GDDR6 memory.

to train the whole model, we used Curated Colon Disease and Kvasir v2 datasets over 80 epochs using Adam optimizer, the binary cross entropy loss function, a batch size of 32

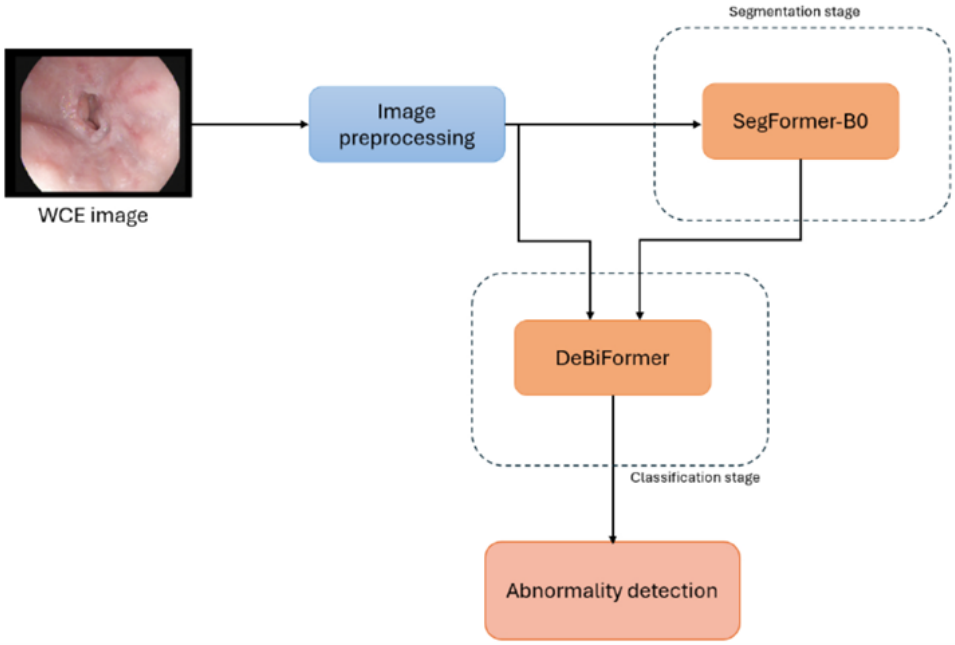

Fig. 4. Overview of the proposed method.

and an initial learning rate $\alpha = 0.0002$. The proposed model was implemented in PyTorch and trained using an NVIDIA RTX A4000 GPU with 16 GB of memory. The computational resources were provisioned through the Paperspace Gradient platform. We utilized a pre-configured GPU-optimized container, which included PyTorch 1.12, CUDA 11.6, and cuDNN 8.

D. Results

In this section, the potential of the proposed method is shown. For this purpose, comprehensive experiments are conducted based on confusion matrix analysis, which collectively offer insights into the model's performances. A confusion matrix is a tabular representation where the model's predicted class labels are compared with the ground truth labels. The matrix is structured to provide counts of four fundamental outcomes for each class:

- True Positive (TP): The model correctly predicted the positive class. (Reality: Yes, Prediction: Yes)
- True Negative (TN): The model correctly predicted the negative class. (Reality: No, Prediction: No)
- False Positive (FP): The model incorrectly predicted the positive class when it was actually negative. Also known as a Type I error. (Reality: No, Prediction: Yes)
- False Negative (FN): The model incorrectly predicted the negative class when it was actually positive. Also known as a Type II error. (Reality: Yes, Prediction: No)

Based on these basic metrics, a total of four specific performance metrics were used : accuracy, precision, specificity and recall.

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Instances}} \quad (1)$$

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \quad (2)$$

$$\text{Specificity} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}} \quad (3)$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \quad (4)$$

After conducting preliminary pre-training, the SegFormer-B0 model attains accuracy values of 98.80% and 91.19%, and loss values of 0.03 and 0.14 on the training and validation sets, respectively. Thus, SegFormer-B0 shows a high potential in improving the performance of segmentation in GI abnormalities detection as can be seen in Fig. 5.

In our experiments, we also compared the proposed method with four deep learning models including Resnet50 [26], VGG16 [27], InceptionV3 [28] and Vit-S16 [29]. To guarantee a scientific and logical comparison of methods, all experiments were carried out in the same environment. The comparison results are shown in Tables III and IV.

On the Curated Colon Disease dataset, our methods achieves the highest accuracy of 99.30%, followed by VGG16 (93.06%) and Vit-S16 (91.83%). Our method attains also the highest values across all other metrics reaching 99.30% precision, 99.30% recall and 99.77% specificity.

The proposed method surpasses all the four methods also On the Kvasir v2 dataset, achieving an accuracy of 94.13%

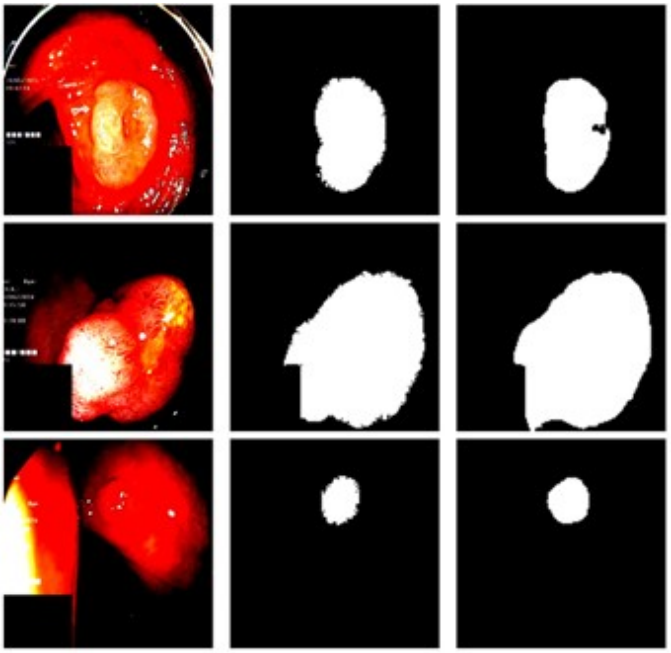

Fig. 5. Visual results of segmentation using SegFormer-B0 model. From left to right: Input image, Ground truth, SegFormer-B0 model.

TABLE III. QUANTITATIVE COMPARISON BETWEEN OUR METHOD AND DIFFERENT MODELS ON THE CURATED COLON DISEASE DATASET

Methods	Metrics			
	Accuracy (%)	Precision (%)	Recall (%)	Specificity (%)
Resnet50	83.40	84.01	83	94.40
VGG16	93.06	93	90.47	97.69
InceptionV3	74.09	83.46	62.22	91.36
Vit-S16	91.83	91.98	91.73	97.33
Our method	**99.30**	**99.30**	**99.30**	**99.77**

accuracy, 94.18% precision, 94.12% recall and 99.16% specificity.

Overall, compared against four prominent deep learning models, our model achieves strong performance on both Curated Colon Disease and Kvasir v2 datasets, underscoring the efficacy and the potential of our approach.

V. CONCLUSION

In this work, we developed and evaluated a novel, two-stage deep learning framework for the automated segmentation, and classification of abnormalities in Wireless Capsule Endoscopy images. The methodology leverages the complementary strengths of two state-of-the-art Transformer-based architectures: SegFormer-B0 for precise pixel-level segmentation of potential pathological regions, and DeBiFormer for the accurate classification of these segmented regions into specific abnormality types by means of its flexible Deformable Bi-level Routing Attention module. Segmentation provides precise spatial localization that can inform and improve classification accuracy, while classification results can guide and refine segmentation boundaries. The experiments shows that our unified framework that simultaneously perform both segmentation and classification significantly enhances performances in WCE image analysis. A comparative studies was conducted to validate the effectiveness of our approach. The results show that our approach achieves the highest values in terms of accuracy, precision, recall and specificity outperforming four common deep learning methods : Resnet50, VGG16, Vit-S16 and InceptionV3.

TABLE IV. QUANTITATIVE COMPARISON BETWEEN OUR METHOD AND DIFFERENT MODELS ON THE KVASIR V2 DATASET

Methods	Metrics			
	Accuracy (%)	Precision (%)	Recall (%)	Specificity (%)
Resnet50	83.40	84.01	0.83	94.40
VGG16	93.06	93	90.47	97.69
InceptionV3	74.09	83.46	62.22	91.36
Vit-S16	91.83	91.98	91.73	97.33
Our method	**94.13**	**94.18**	**94.12**	**99.16**

While our approach generally succeeds in identifying the most important anomalies, it is less efficient in the presence of less prominent anomalies and noise such as air bubbles and food debris. In future work, we aim to address these limitations and improve our model performance by including transfer Learning techniques and extending our evaluation to other public datasets offering additional variability in WCE images.

REFERENCES

[1] P. Collins, "Video-capsule endoscopy – current and future role," *Medicine*, vol. 52, no. 5, pp. 317–319, May 2024.

[2] Z. Xiao and L. N. Feng, "A Study on Wireless Capsule Endoscopy for Small Intestinal Lesions Detection Based on Deep Learning Target Detection," *IEEE Access*, vol. 8, pp. 159 017–159 026, 2020.

[3] V. Sadeghi, Y. Sanahmadi, M. Behdad, A. Vard, M. Sharifi, A. Raeisi, M. Nikkhah, and A. Mehridehnavi, "Pixel-wise annotation for clear and contaminated regions segmentation in wireless capsule endoscopy images: A multicentre database," *Data in Brief*, vol. 57, p. 110927, Dec. 2024.

[4] S. Soffer, E. Klang, O. Shimon, N. Nachmias, R. Eliakim, S. Ben-Horin, U. Kopylov, and Y. Barash, "Deep learning for wireless capsule endoscopy: A systematic review and meta-analysis," *Gastrointestinal Endoscopy*, vol. 92, no. 4, pp. 831–839.e8, Oct. 2020.

[5] T. T. Habe, K. Haataja, and P. Toivanen, "Review of Deep Learning Performance in Wireless Capsule Endoscopy Images for GI Disease Classification," *F1000Research*, vol. 13, p. 201, 2024.

[6] S. Bandopadhyay, J. R. Chowdhury, R. Shaw, S. Paul, and S. Banerjee, "Vision Transformer-Based Classification of Gastrointestinal Ulcers Using WCE Images," in *Computational Intelligence in Pattern Recognition*, A. K. Das, J. Nayak, B. Naik, M. Himabindu, S. Vimal, and D. Pelusi, Eds. Singapore: Springer Nature, 2025, pp. 367–378.

[7] A. Garbaz, S. Lafraxo, S. Charfi, M. El Ansari, L. Koutti, and M. Salihoun, "Combined deep convolutional neural networks for abnormality classification in wireless capsule endoscopy images," *Multimedia Tools and Applications*, Mar. 2025.

[8] X. Li, Q. Wu, and K. Wu, "Wireless capsule endoscopy anomaly classification via dynamic multi-task learning," *Biomedical Signal Processing and Control*, vol. 100, p. 107081, Feb. 2025.

[9] P. K. Muruganantham and S. M. Balakrishnan, "Uncertainty-driven active learning in a deep semi-supervised framework for WCE image classification," *Results in Engineering*, vol. 27, p. 106174, Sep. 2025.

[10] M. Bordbar, M. S. Helfroush, H. Danyali, and F. Ejtehadi, "Wireless capsule endoscopy multiclass classification using three-dimensional deep convolutional neural network model," *BioMedical Engineering OnLine*, vol. 22, no. 1, p. 124, Dec. 2023.

[11] M. Madhura Prakash and G. N. Krishnamurthy, "An innovative transfer learning-polyp detection from wireless capsule endoscopy videos with optimal key frame selection and depth estimation," *Biomedical Signal Processing and Control*, vol. 108, p. 107963, Oct. 2025.

[12] Z. Amiri, H. Hassanpour, and A. Beghdadi, "Combining deep features and hand-crafted features for abnormality detection in WCE images," *Multimedia Tools and Applications*, vol. 83, no. 2, pp. 5837–5870, Jan. 2024.

[13] A. Caroppo, P. Siciliano, and A. Leone, "An expert system for lesion detection in wireless capsule endoscopy using transfer learning," *Procedia Computer Science*, vol. 219, pp. 1136–1144, Jan. 2023.

[14] D. He, R. Zhang, Y. Liang, J. Chen, and Y. Li, "Multi-View Attention Network With Iterative Feature Refinement and Boundary Awareness for Endoscopic Image Segmentation," *IEEE Access*, vol. 13, pp. 131 217–131 233, 2025.

[15] S. Charfi, M. E. Ansari, L. Koutti, A. Ellahyani, and I. Eljaafari, "Modified residual attention network for abnormalities segmentation and detection in WCE images," *Soft Computing*, vol. 28, no. 9, pp. 6923–6936, May 2024.

[16] A. Quindós, P. Laiz, J. Vitrià, and S. Seguí, "Self-supervised out-of-distribution detection in wireless capsule endoscopy images," *Artificial Intelligence in Medicine*, vol. 143, p. 102606, Sep. 2023.

[17] S. Fan, S. Wei, K. Wei, S. Liu, Y. Fan, B. Lv, and L. Li, "A novel automatic locating method for pylorus and ileocecal valve in wireless capsule endoscopy," *Biomedical Signal Processing and Control*, vol. 100, p. 106969, Feb. 2025.

[18] R. Fernandes, A. Pessoa, J. Nogueira, A. Paiva, I. Paçal, M. Salgado, and A. Cunha, "Evaluation of Deep Learning Models in Search by Example using Capsule Endoscopy Images," *Procedia Computer Science*, vol. 239, pp. 2065–2073, Jan. 2024.

[19] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers," Oct. 2021.

[20] N. H. B. Long, C. Zhang, Y. Shi, T. Hirakawa, T. Yamashita, T. Matsui, and H. Fujiyoshi, "DeBiFormer: Vision Transformer with Deformable Agent Bi-level Routing Attention," Oct. 2024.

[21] J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, and F. Vilariño, "WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians," *Computerized Medical Imaging and Graphics: The Official Journal of the Computerized Medical Imaging Society*, vol. 43, pp. 99–111, Jul. 2015.

[22] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. de Lange, D. Johansen, and H. D. Johansen, "Kvasir-SEG: A Segmented Polyp Dataset," Nov. 2019.

[23] F. J. P. Montalbo, "Diagnosing gastrointestinal diseases from endoscopy images through a multi-fused CNN with auxiliary layers, alpha dropouts, and a fusion residual block," *Biomedical Signal Processing and Control*, vol. 76, p. 103683, Jul. 2022.

[24] K. Pogorelov, K. R. Randel, C. Griwodz, S. L. Eskeland, T. de Lange, D. Johansen, C. Spampinato, D.-T. Dang-Nguyen, M. Lux, P. T. Schmidt, M. Riegler, and P. Halvorsen, "KVASIR: A Multi-Class Image Dataset for Computer Aided Gastrointestinal Disease Detection," in *Proceedings of the 8th ACM on Multimedia Systems Conference*, ser. MMSys'17. New York, NY, USA: Association for Computing Machinery, Jun. 2017, pp. 164–169.

[25] Y. E. Khalfaoui, B. Alibouch, and A. F. E. Ouafdi, "Combining Deep CNNs and Superpixels for Polyp Segmentation in Colonoscopy Images," in *2023 10th International Conference on Wireless Networks and Mobile Communications (WINCOM)*, Oct. 2023, pp. 1–6.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," Dec. 2015.

[27] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," Apr. 2015.

[28] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," Dec. 2015.

[29] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," Jun. 2021.