Automated Segmentation of Colorectal Cancer on Routine Computed Tomography Using Deep Learning

Yu-Xin Lin
R12631070@ntu.edu.tw

ABSTRACT

Colorectal cancer (CRC) is the third most common cancer globally, with significant morbidity and mortality rates. While colonoscopy remains the gold standard for CRC detection and prevention, routine computed tomography (CT) scans can serve as an alternative for certain patient populations despite their lower diagnostic accuracy. This study addresses the challenge of segmenting colorectal cancer tumors in routine CT images by evaluating the performance of traditional segmentation models (nnU-Net and Swin-UMamba) against a prompt-driven model (Promise). Utilizing both public and internal datasets, we demonstrate that the Promise model significantly outperforms traditional approaches across all evaluation metrics, largely due to its prompt-assistance mechanism, which minimizes incorrect predictions in irrelevant regions. Among traditional models, nnU-Net exhibits superior reliability and stability. This study compares prompt-driven and traditional segmentation models for CRC tumor segmentation, providing empirical evidence of the advantages of prompt-driven models in medical image analysis. Our findings highlight the potential of these models to reduce missed diagnoses, alleviate radiologists' workloads, and improve diagnostic consistency.

The code and model weights mentioned in this paper can be found at:
https://www.space.ntu.edu.tw/navigate/a/#/s/4457165EAFE04A09A69EB96E61F6C2AF6BL

Keywords: AI, deep learning, colon cancer, segmentation

1. INTRODUCTION

Colorectal cancer (CRC) is the third most common cancer worldwide and results in significant morbidity and mortality. Colonoscopy is the most commonly used and efficient diagnostic method for CRC. Combining colonoscopy with polypectomy can lower the incidence and mortality rates of CRC.

An alternative for patients who are unable to tolerate colonoscopy is computed tomographic (CT) virtual colonoscopy, which is comparable with colonoscopy in detecting CRC and polyps larger than 10mm. Nevertheless, CT virtual colonography necessitates bowel preparation and insufflation of gas, which might be difficult for certain patient groups, such as the elderly or frail. Previous studies have investigated the value of routine, unprepared abdominopelvic CT for the detection of CRC and shown various performances depending on different study designs.

Although the accuracy of routine CT is not as good as that of colonoscopy and CT virtual colonography, it is still able to detect a considerable number of CRC, and has better availability and patient compliance. However, the miss rate of routine CT was significant. A study reported that 41% of the routine abdominopelvic CTs before the diagnosis of CRC had the findings undetected by radiologists.

Suppose the CT scans are performed for reasons other than the detection of colorectal lesions, such as routine follow-up of abdominal aortic aneurysm or hepatocellular carcinoma. In that case, radiologists may fail to detect the incidental CRC because of perceptual errors. A delay in the diagnosis of CRC on a CT scan may worsen the patient's survival.

Artificial intelligence may have the potential to reduce perceptual and interpretive errors in diagnostic radiology. Numerous deep-learning models have been proposed to detect colorectal polyps or cancer. Most of them were based on colonoscopy or CT virtual colonography.

Furthermore, training radiologists to be able to identify colorectal cancer tumors in traditional computed tomography scans is far more difficult than looking at other organs and requires years of specialized training. In addition, because tumors do not have clear borders on traditional CT scans, the results marked by different doctors can vary significantly.

Therefore, this study aims to develop such a model. It can reduce the missed diagnosis rate, reduce the workload of doctors, reduce the training costs of doctors and improve consistency.

2. METHODOLOGY

2.1 Data Source

The training and testing data are sourced from the Medical Segmentation Decathlon (MSD) dataset [1] and the National Taiwan University Hospital (NTUH) internal dataset , containing 126 and 70 samples, respectively. The MSD dataset is a generalizable 3D Semantic Segmentation dataset that contains 10 tasks. In our research, we used Task 10 Colon Cancer subset.

2.2 Preprocessing

File format: Common medical image formats include DICOM (suffix .dcm), MHD (suffix .mhd and .raw), NIFTI (suffix .nii or .nii.gz), and MRB (suffix .mrb). Among them, NIFTI files are the most common in deep learning, but hospitals are usually accustomed to using DICOM or MRB. We need to convert to NIFTI files before subsequent operations. In the nnU-Net framework, additional programs must convert the training data into numpy arrays (.npz) and corresponding pickle files (.pkl). The numpy files store each image's voxel intensity, and the pickle files store the image spacing, direction, and origin information. However, the data used for testing does not need to undergo this step and only needs to maintain NIFTI.

Adjust Intensity: In CT images, the intensity of the image is represented by Hounsfield unit (HU) instead of the general grayscale 8-bit intensity. HU is the linear attenuation coefficient of the measurement medium and is linearly mapped to the space of HU_{water} 0 and HU_{air} -1000. Usually the range may be from -1024 to 2048 etc. When observing abdominal CT, it is usually customary to set window 400 and level 40 to facilitate observation of the organs in the abdomen.

Crop: In 3D medical imaging, a significant portion of the volume often consists of irrelevant background regions, and the size of these images can be exceptionally large. This can lead to excessive computational resource consumption during training. To improve computational efficiency, nnU-Net automatically detects the foreground region of an image and applies bounding boxes to crop out relevant areas. This process reduces the burden of processing irrelevant background data, enabling the model to focus on meaningful information while lowering hardware requirements.

Resample: Due to variations in patient anatomy and machine settings, CT images often have inconsistent voxel spacing, origins, and other metadata. To address this, resampling is performed to standardize voxel spacing across all images. nnU-Net employs a data-driven strategy by analyzing the spacing distribution across the dataset. For datasets with relatively uniform spacing, the median spacing is selected as the target; for highly anisotropic datasets (where the maximum spacing is more than three times the minimum spacing), the 10th percentile spacing is chosen instead. During resampling, trilinear interpolation is applied to the image data, while nearest-neighbor interpolation is used for the labels to preserve segmentation accuracy.

Normalization: For CT images, nnU-Net applies Z-score normalization to standardize voxel intensities. First, the framework calculates the intensity distribution within the foreground region across the entire training set, trimming outliers by retaining only the range between the 0.5th and 99.5th percentiles to avoid the influence of extreme values. Then, the mean and standard deviation of the remaining values are computed. Each voxel value is subsequently standardized by subtracting the mean and dividing by the standard deviation, resulting in normalized Z-scores. This ensures robustness during the training process.

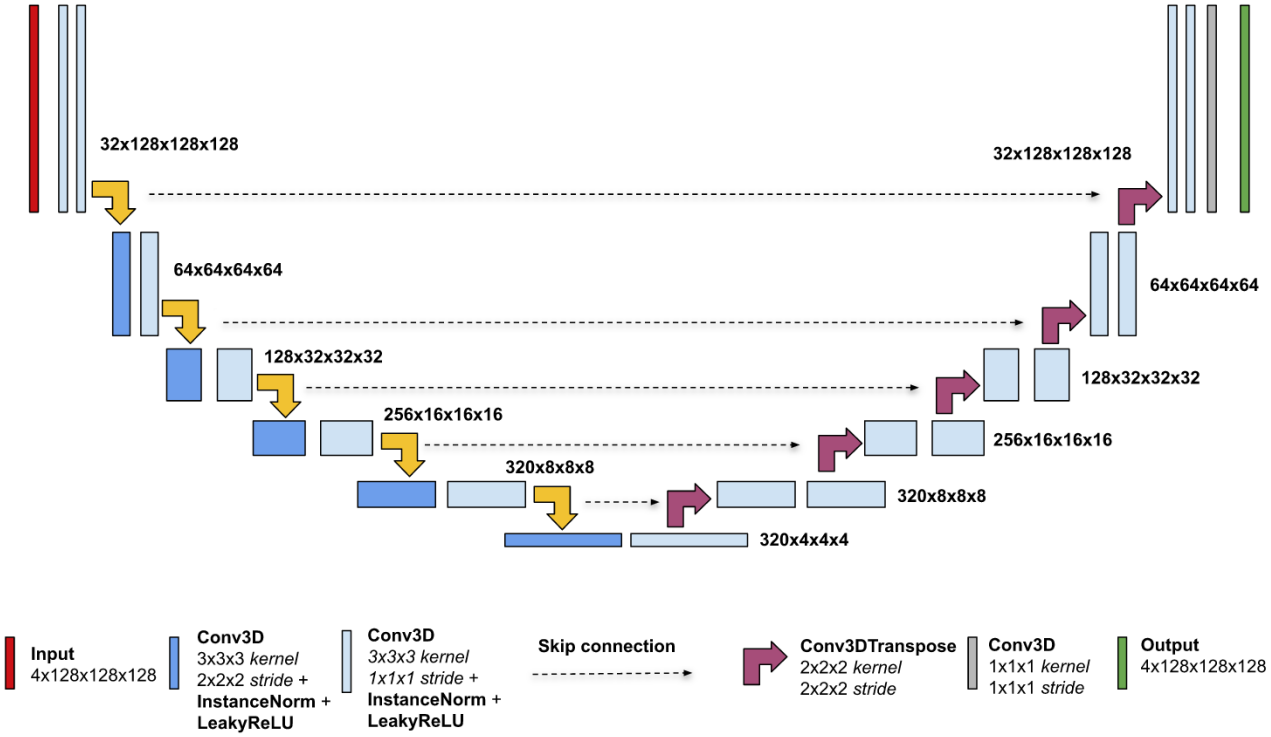

Figure 1. The encoder transforms input volume until it reaches a size of 2 × 2 × 2 in the bottleneck. Then the decoder upsamples it with transposed convolutions back to the original input shape of 128 × 128 × 128. Additional two output heads are used for deep supervision loss.

5 Fold Cross Validation: During preprocessing, nnU-Net employs a 5-fold cross-validation strategy to partition the training dataset into five distinct combinations of training and validation sets. This approach enhances the model's stability and generalization capabilities, effectively mitigating biases caused by imbalanced data distributions. In the inference phase, nnU-Net averages the voxel-wise probabilities predicted by all five models (ensemble learning), further improving the accuracy and reliability of segmentation results.

2.3 Traditional Segment Models

nnU-Net [2] is a highly adaptive medical image segmentation framework designed to address the challenges posed by the diversity of medical imaging datasets. Unlike traditional deep learning models, nnU-Net automatically tailors a wide range of operations—including preprocessing, network architecture adjustments, and hyperparameter optimization—based on the characteristics of each dataset, delivering efficient segmentation performance.

Built upon the standard U-Net architecture, nnU-Net incorporates multilevel feature extraction, advanced data augmentation strategies, and robust, adaptive loss functions. This approach has enabled nnU-Net to achieve top rankings across various medical image segmentation benchmarks (e.g., task-based challenges), surpassing many customized state-of-the-art models. Its design principles of generality and scalability have made nnU-Net a widely adopted baseline in medical image segmentation research, serving as an indispensable tool in both academic and industrial settings.

Swin-UMamba [3] is an innovative medical image segmentation model that combines the architectural strengths of Swin Transformer and U-Net. Its core innovation lies in further optimizing the backbone network of the Swin Transformer by replacing it with the novel VMamba architecture and integrating U-Net's skip connection mechanism. This design effectively leverages both local and global feature representations.

Swin-UMamba is tailored to the unique characteristics of medical imaging data, enhancing the model's ability to capture high-resolution and structural features. The VMamba [4] architecture excels at handling multiscale information, demonstrating notable advantages in segmenting small lesions and addressing fuzzy boundaries. Meanwhile, skip connections ensure that the model preserves fine-grained details while extracting deep features, resulting in more precise segmentation outcomes.

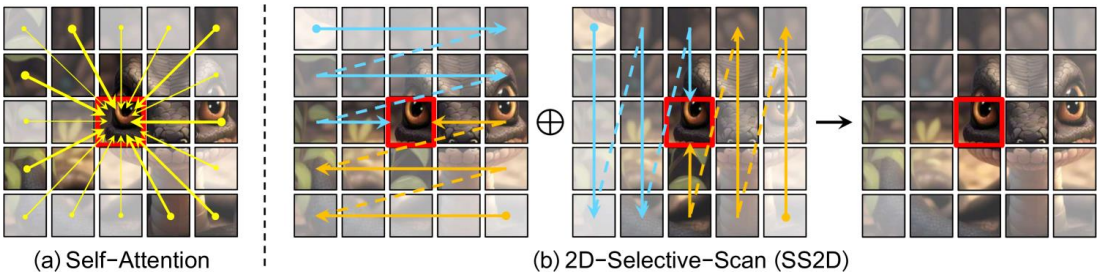

Figure 2. Comparison of correlation establishment between image patches via (a) self-attention and (b) the proposed 2D-Selective-Scan (SS2D). Red boxes indicate the query image patch, with patch opacity representing the degree of information loss.

2.4 Prompt Driven Segment Model

Promise [5] is an innovative model designed for 3D medical image segmentation tasks. Its core concept builds upon Meta's Segmentation Anything Model (SAM) [6], employing a prompt-driven learning approach that allows users to guide the segmentation process through predefined prompts, such as organ locations, anatomical structure characteristics, or lesion labels. Unlike traditional segmentation models, Promise does not require explicit tumor localization during training; instead, it directly segments the tumor based on the given prompt.

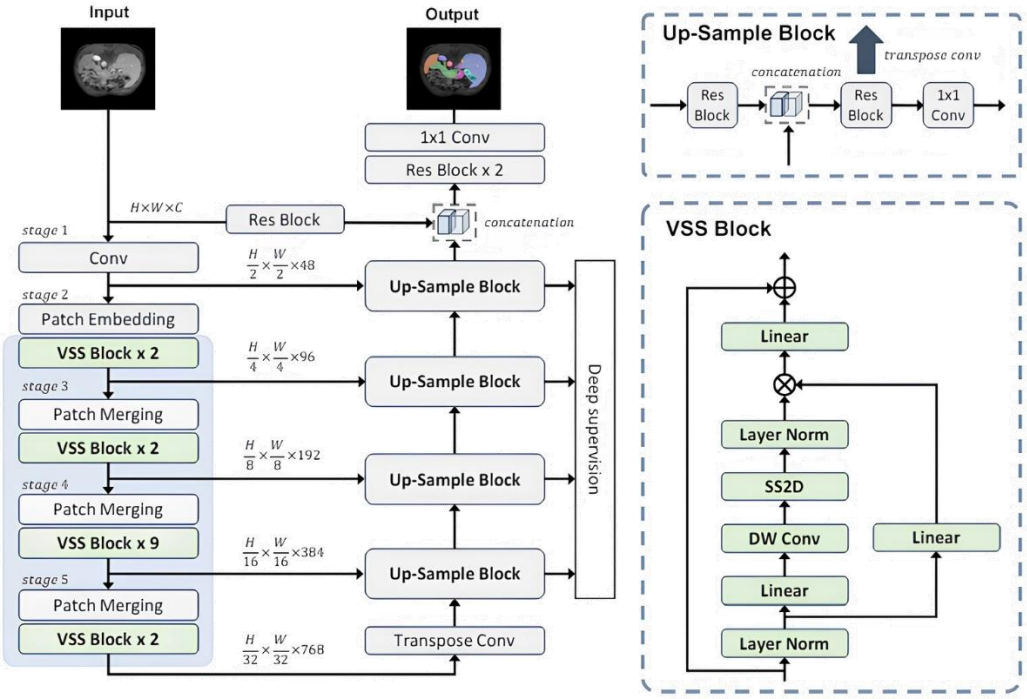

Figure 3. The overall architecture of Swin-UMamba. Swin-UMamba can leverage the power of vision foundation models by loading the weights of pretrained models. Each block within the blue box was initialized with the ImageNet pretrained weights.

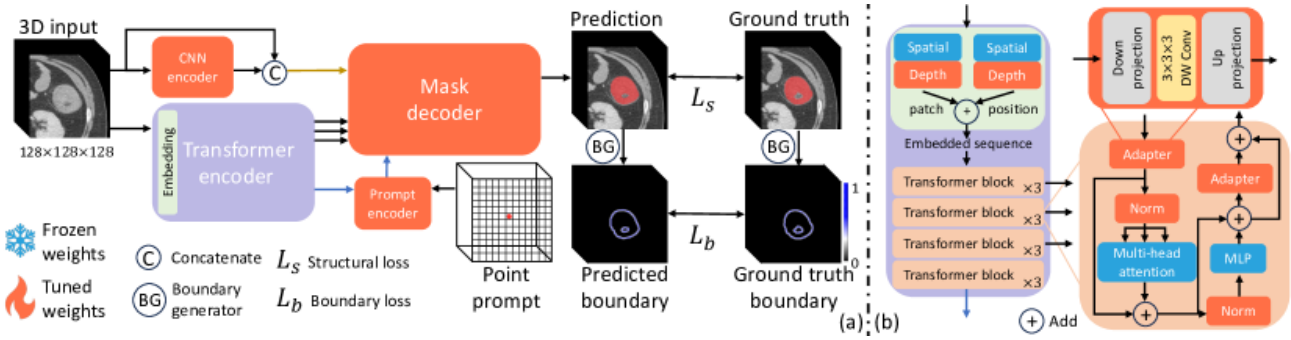

Figure 4. The proposed framework (ProMISe) and details of transformer encoder are shown as (a) and (b), respectively.

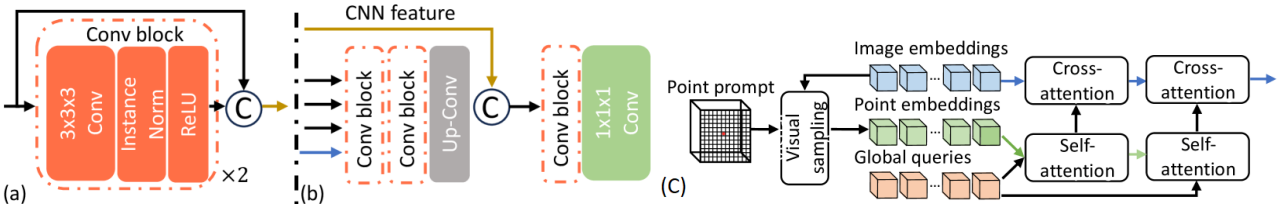

Figure 5. The details of (a) CNN encoder, (b) mask decoder, (c) and the proposed prompt encoder.

Promise leverages a powerful 3D deep neural network for feature extraction, integrating a multiscale attention mechanism to capture fine-grained structures and global context with precision. Additionally, its built-in dynamic network adjustment module adapts the architecture dynamically based on the input prompts, enhancing the model's adaptability and segmentation accuracy.

Promise's design takes into account the specific challenges of medical imaging, including high resolution, low contrast, and irregular lesion shapes. Experimental results have demonstrated its exceptional performance across various publicly available 3D medical imaging datasets, such as those for tumor and organ segmentation. Notably, Promise has shown outstanding practical value in handling rare lesions or limited annotated data, marking a significant breakthrough in the application of prompt-driven learning to medical image segmentation.

2.5 Hardware Environment

For this study, we used a server with an Intel(R) Xeon(R) Silver 4110 CPU @ 2.10GHz, running CentOS Linux release 7.9.2009 (Core), and equipped with 2 NVIDIA RTX 3090 graphics cards (24GB).

3. RESULTS

Based on the experimental results, we conducted a comprehensive comparison of the performance of the three models on the task of colorectal cancer tumor segmentation. The table presents four key metrics: Dice coefficient, Normalized Surface Distance (NSD, δ=1), 95% Hausdorff Distance (HD95), and Recall.

The Promise model outperformed all others across all evaluation metrics. Its Dice coefficient reached 0.64789, significantly higher than nnU-Net's 0.536395 and Swin-UMamba's 0.32465, indicating a clear advantage in segmentation accuracy. For the NSD metric, Promise achieved a score of 0.557895, also surpassing the other two models, reflecting its high similarity to the ground truth labels.

Particularly noteworthy is the HD95 metric, where Promise recorded a value of only 16.1817, markedly lower than nnU-Net's 124.4473 and Swin-UMamba's 92.66501. This demonstrates a significant improvement in boundary accuracy,

substantially reducing segmentation errors. In terms of recall, Promise achieved 0.77604, notably higher than the other two models, indicating its superior ability to identify and include tumor regions effectively.

Among the traditional segmentation models, nnU-Net outperformed Swin-UMamba overall. nnU-Net exhibited superior performance in Dice coefficient (0.536395 vs. 0.32465), NSD (0.44307 vs. 0.283745), and recall (0.62459 vs. 0.29756). The only exception was the HD95 metric, where Swin-UMamba (92.66501) slightly outperformed nnU-Net (124.4473).

Overall, these results clearly highlight the superiority of the prompt-driven Promise model in colorectal cancer tumor segmentation tasks. It excels not only in segmentation accuracy and boundary localization but also in the ability to effectively identify tumor regions. Among the traditional models, while nnU-Net does not match the performance of Promise, it still outperforms Swin-UMamba in most metrics, demonstrating its reliability as a mature segmentation model.

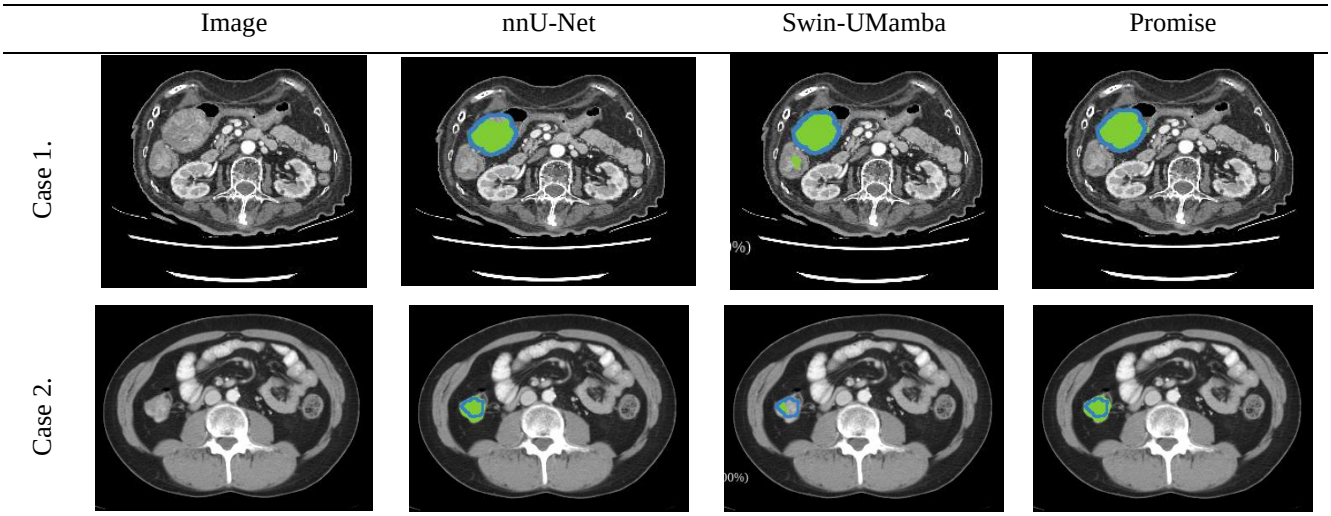

Figure 6. Slices of the original computed tomography image and the test results of each model on the test set. The real area is represented by the hollow blue area, and the inferred area is represented by the solid green area.

Table 1. Compares three segmentation models for colorectal cancer tumor detection. The Promise model demonstrates superior performance across all metrics, with nnU-Net showing moderate performance and Swin-UMamba achieving the lowest scores overall.

	Dice	NSD (δ=1)	HD95	Recall
nnU-Net	0.536395	0.44307	124.4473	0.62459
Swin-UMamba	0.32465	0.283745	92.66501	0.29756
Promise	0.64789	0.557895	16.1817	0.77604

4. CONCLUSIONS.

This study addresses the challenge of segmenting colorectal cancer tumors in routine computed tomography (CT) images by comparing the performance of two traditional image segmentation models (nnU-Net and Swin-UMamba) with an emerging prompt-driven model (Promise). Using both a public dataset and an internal dataset for model training and testing, we have drawn the following key conclusions: the prompt-driven model Promise significantly outperforms traditional segmentation models across all evaluation metrics, primarily due to its prompt-assistance mechanism, which effectively prevents incorrect predictions in irrelevant regions. Among traditional segmentation models, nnU-Net

demonstrates superior performance in most metrics, showcasing its stability and reliability in medical image segmentation tasks. These findings underscore the importance of selecting appropriate segmentation models to enhance the accuracy of colorectal cancer tumor detection.

The primary contribution of this study lies in being the first to compare the effectiveness of a prompt-driven model with traditional segmentation models in the context of colorectal cancer tumor segmentation. This provides valuable empirical evidence for the field of medical image analysis, highlighting the superiority of prompt-driven models for specific tasks. Furthermore, these results offer a basis for selecting suitable tumor segmentation tools in clinical practice, which could reduce missed diagnoses and alleviate the workload of healthcare professionals.

Building upon these findings, we recommend several directions for future research. First, exploring the integration of prompt-driven models with traditional segmentation models could leverage their respective strengths. Second, validating the performance of these models on larger clinical datasets is crucial to assess their potential for real-world medical applications. Third, investigating the applicability of these models to other medical imaging modalities, such as magnetic resonance imaging (MRI), could extend their scope. Additionally, further optimization of the prompt design in the Promise model to enhance its generalizability across different tumor segmentation tasks is a promising avenue for exploration.

Overall, this study provides new insights into the automated segmentation of colorectal cancer tumors, paving the way for improved diagnostic accuracy and efficiency. As technological advancements continue, we anticipate that these models will play a more significant role in clinical practice, ultimately enhancing patient care and treatment outcomes. Future research should further explore the applications of artificial intelligence in medical image analysis to drive the development of precision medicine.

REFERENCE

[1] M. Antonelli, A. Reinke, S. Bakas *et al.*, “The Medical Segmentation Decathlon,” Nature Communications, 13, (2021).

[2] F. Isensee, J. Petersen, A. Klein *et al.*, “nnU-Net: Self-adapting Framework for U-Net-Based Medical Image Segmentation,” ArXiv, abs/1809.10486, (2018).

[3] J. Liu, H. Yang, H.-Y. Zhou *et al.*, "Swin-UMamba: Mamba-based UNet with ImageNet-based pretraining."

[4] Y. Liu, Y. Tian, Y. Zhao *et al.*, “VMamba: Visual State Space Model,” ArXiv, abs/2401.10166, (2024).

[5] H. Li, H. Liu, D. Hu *et al.*, “Promise: Prompt-Driven 3D Medical Image Segmentation Using Pretrained Image Foundation Models,” 2024 IEEE International Symposium on Biomedical Imaging (ISBI), 1-5 (2023).

[6] A. Kirillov, E. Mintun, N. Ravi *et al.*, “Segment Anything,” 2023 IEEE/CVF International Conference on Computer Vision (ICCV), 3992-4003 (2023).