UltraLight Med-Vision Mamba for Classification of Neoplastic Progression in Tubular Adenomas

Aqsa Sultana*, Nordin Abouzahra*, Ahmed Rahu, MD†, Brian Shula‡,
Brandon Combs§, Derrick Forchetti, MD§, Theus Aspiras, PhD*, Vijayan K. Asari, PhD*
*Dept. of Electrical and Computer Engineering, University of Dayton, Dayton, OH USA
† Dept. of Pathology, University of Toledo Medical Center, Toledo, OH, USA
‡ Lead Mechanical Engineer, Honeywell International Inc., South Bend, IN, USA
§ Dept. of Pathology, South Bend Medical Foundation, South Bend, IN, USA

***Abstract*—Identification of precancerous polyps during routine colonoscopy screenings is vital for their excision, lowering the risk of developing colorectal cancer. Advanced deep learning algorithms enable precise adenoma classification and stratification, improving risk assessment accuracy and enabling personalized surveillance protocols that optimize patient outcomes. UltraLight Med-Vision Mamba, a state-space-based model (SSM), has excelled in modeling long- and short-range dependencies and image generalization, critical factors for analyzing whole slide images. Furthermore, UltraLight Med-Vision Mamba's efficient architecture offers advantages in both computational speed and scalability, making it a promising tool for real-time clinical deployment.**

***Index Terms*—Vision Mamba, state space models, medical image classification, biomedical, adenomas, cancer risk**

I. Introduction

Colorectal cancer (CRC) is a major global health challenge; in the United States, it's the third most common cause of cancer and is the second leading cause of cancer-related death [1]. CRC frequently originates from colonic polyps, which are raised protrusions of colonic mucosa of epithelial origin, broadly categorized as adenomatous and serrated. Adenomatous polyps are due to neoplastic proliferation of glands and are a well-established, frequent precursor lesion to CRC [2]. For decades, CRC has been thought to arise through a traditional adenoma-carcinoma pathway [3], and screening guidelines have been established to detect precancerous lesions. The US Preventive Services Task Force recommendations show a substantial benefit to screening asymptomatic individuals starting at age 50 and a moderate benefit starting at age 45 [4]. Tubular adenomas, a subtype within the adenomatous polyps category, are one of such precancerous lesions, and are the primary focus of this discussion.

Tubular adenomas can be classified into two categories: those having low-grade dysplasia and those having high-grade dysplasia. The presence of high-grade dysplasia is a known risk factor for the development of CRC [5] and represents an advanced stage along the adenoma-carcinoma continuum. Despite prophylactic screening efforts, accurately assessing the malignant potential of low-grade tubular adenomas remains a clinical challenge. Traditional histopathological examination is based solely on visual assessment; the naked eye can face limitations in identifying subtle morphological features associated with patients' long-term risks of developing subsequent CRC.

The advent of digital pathology has enabled the generation of high-resolution whole slide images (WSIs). Combining WSIs with advancements in artificial intelligence (AI) has led to powerful new tools augmenting diagnostic accuracy and efficiency in pathology. This synergy between digital pathology and AI promises to improve the sensitivity of risk stratification and other aspects of clinical care that are impossible with traditional light microscopic examination alone [6].

Bridging digital pathology alongside breakthroughs in deep learning offers unprecedented opportunities for objective and quantitative histological analysis. Deep learning models can now analyze, extract, and learn complex patterns directly from image data, identifying subtle morphological patterns and features imperceptible through visual assessment [6]. This study implements Vision Mamba, [7], a novel State Space Model (SSM) [8] architecture, to analyze intricate histological patterns within WSIs of low-grade tubular adenomas to identify subtle indicators associated with subsequent colorectal cancer risk. The model also leverages an ultra-light architecture preventing parameter explosion, making it a promising tool for real-time clinical deployment [8].

II. Methodology

The UltraLight Med-Vision Mamba model [7], [9] adopts an architectural framework akin to convolutional neural networks (CNNs), but with a key distinction: instead of relying on convolutional blocks as its primary feature extractors, it employs Parallel Vision Mamba (PVM) layers. The overall architecture comprises six layers, with the number of channels configured as $[8, 16, 24, 32, 48, 64]$ as shown in Fig. 1. The initial three layers utilize standard convolutional blocks to extract shallow features, while the deeper layers (layers 4 through 6) incorporate PVM layers to capture more complex and nuanced features. The extracted features from three convolutional blocks and two PVM layers are fed into SCAB (spatial and channel attention bridge) module. Adaptive average pooling is performed on feature maps acquired from SCAB and the final PVM layer (stage 6) to standardize spatial dimensions of the feature maps. The pooled features are then concatenated to

The authors graciously thank the South Bend Medical Foundation (SBMF) for providing the whole slide images for this work.

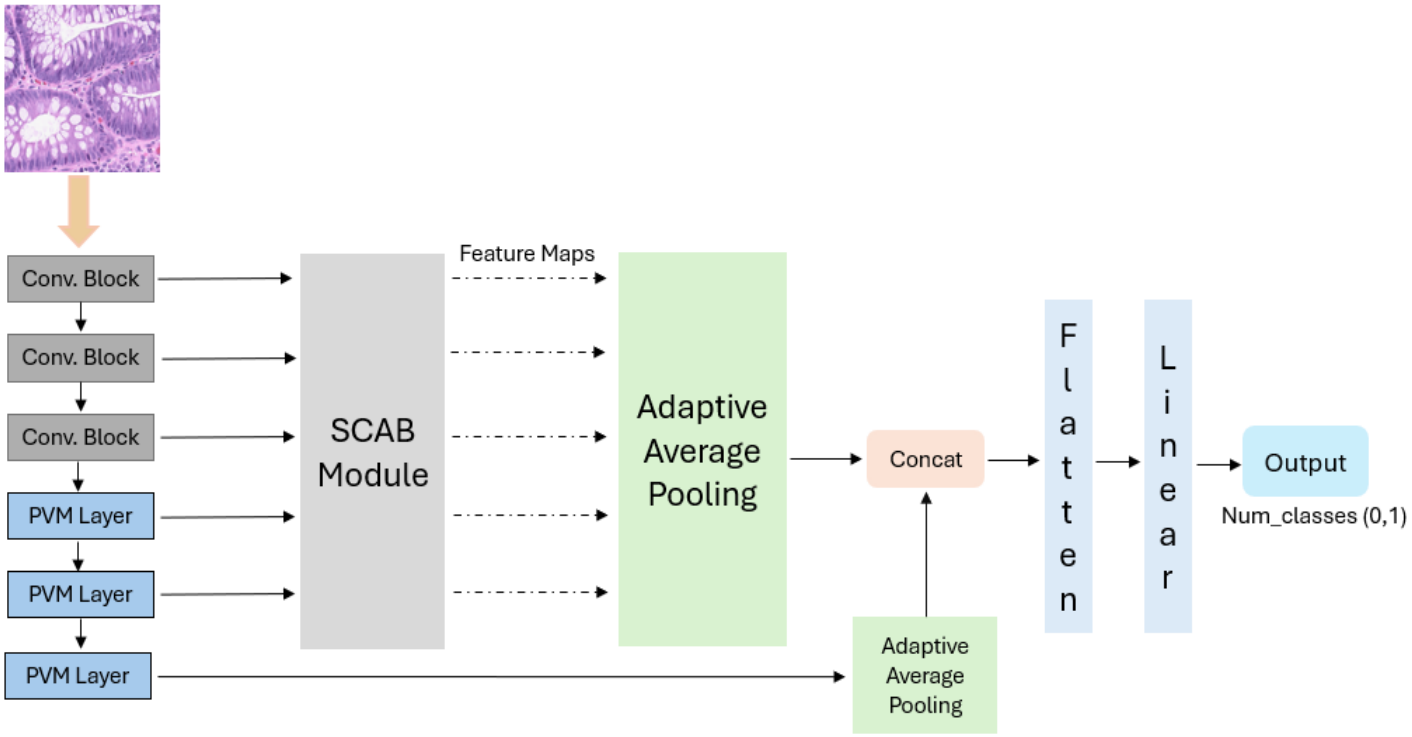


Fig. 1: Architectural structure of UltraLight Med-Vision Mamba model for image classification task.

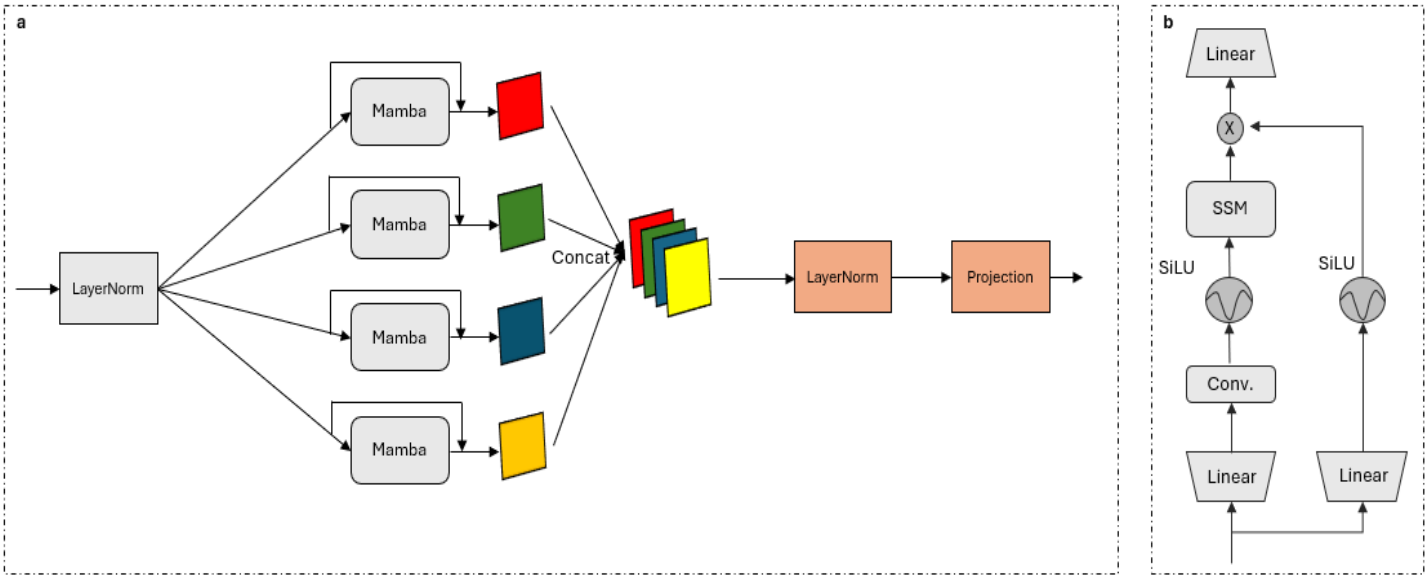


Fig. 2: a) PVM layer in UltraLight Med-Vision Mamba b) Mamba module.

accumulate and retain the relevant information. The classification head first flattens the high-dimensional feature maps into a one-dimensional vector, effectively aggregating the spatially distributed information extracted by the preceding layers. This flattened vector is then passed through fully connected (dense) layers, which serve to map the learned feature representations to the final output space. In this case of colorectal adenoma classification, this typically corresponds to a set of class probabilities, enabling the model to assign a likelihood score to each potential category of control and cancer group.

The Parallel Vision Mamba (PVM) layer [7], [9], as shown in Fig. 2 (a), is also known as the PVM module. It incorporates Mamba blocks with residual connections from the input to the output of Mamba blocks to enhance the model's ability to capture complex spatial relationships. The input first undergoes layer normalization, after which the feature maps are split into four distinct branches, each with a designated number of channels. These branches are processed independently through the Mamba mechanism. The outputs from the Mamba blocks are then combined with residual connections from the original inputs, along with an adjustment factor to optimize learning. The resulting feature maps are concatenated to form four unified feature maps with specific channel dimensions. These concatenated outputs are subsequently normalized again and passed through a projection layer. By processing features in parallel across multiple branches, the PVM module is able to extract multiscale and intricate feature representations using varying kernel sizes. Moreover, this design efficiently reduces the number of parameters by preserving the same receptive field, thereby mitigating the parameter growth typically associated with increasing channel dimensions–an important consideration, as the parameter count in Mamba layers is highly sensitive to input channel size.

The model performance is further improved by the addition of SCAB module, also known as the Spatial and Channel Attention Bridge [7], [10], for feature propagation. Spatial attention bridge consists of max-pooling, average pooling, and extended convolution of shared weights. Channel attention bridge includes fully connected layers (FCL), global average pooling (GAP), concatenation, and sigmoid activation function. The SCAB module enhances the sensitivity, ability of the model to converge, and fusion of multi-scale features of different scales [7], [9].

III. Training and Experimental Results

A. Training Method

The baseline for the experiment was established by training all models—Vision Transformer (ViT), swin Transformer, and UltraLight Med-Vision Mamba—for 100 epochs using Stochastic Gradient Descent (SGD) with a momentum of 0.9 and an initial learning rate of 0.001. UltraLight Med-Vision Mamba was further fine-tuned with a learning rate range of 0.0001 to 0.05 and trained for 300 epochs. Binary Crossentropy was used as the loss function. The training strategy incorporated the OneCycle Learning Rate (OneCycleLR) scheduler and Stochastic Weight Averaging (SWA) to stabilize training. The experiments were implemented using the PyTorch framework in Python on a NVIDIA GeForce RTX Titan GPU.

B. Transformer Based Models

Vision Transformer [11], also known as ViT, divides the input image into fixed-size patches, flattens them, and treats them like tokens in a sequence similar to Natural language Processing (NLP) [12]. The patches are then processed using multi-self-attention-heads to capture global context. While powerful, ViT requires large datasets and high computational resources.

Swin Transformer [13] also known as **S**hifted **Win**dow Transformer is a more efficient variant of Transformers that hierarchically processes images using non-overlapping local windows. It introduces shifting window mechanism that allows cross-window connections while maintaining computational efficiency. This makes the Swin Transformer more scalable for smaller datasets and dense prediction tasks than ViT.

C. Model Parameters

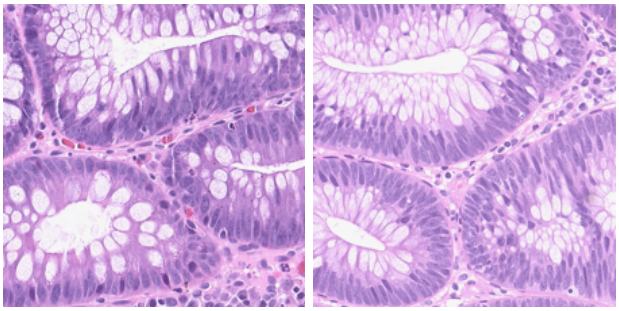
The comparison between the number of model parameters for the three architectures is presented in Table I. ViT has the highest number of parameters at $7,398,785$, indicating large model size and computational overhead. Swin Transformer, with its heirarchical architectural structure, is designed for more efficient computation significantly reduces the model parameter count to $598,099$. Lastly, UltraLight Med-Vision Mamba, dramatically reduces the model parameter to only $49,641$—making it the most lightweight model among the three. This suggests that UltraLight Med-Vision Mamba is highly optimized for efficiency, better trade-off between performance and computational cost, especially for the real-time clinical deployment.

TABLE I: Model parameter comparison for ViT, Swin Transformer and UltraLight Med-Vision Mamba.

Model	Model Parameters
ViT	7,398,785
Swin Transformer	598,099
UltraLight Med-Vision Mamba	**49,641**

D. Dataset

The original whole slide images (WSIs) were tiled at 1024 × 1024 pixel resolution, with 3 color channels and then resized into smaller tiles of 224 × 224 pixels with 3 color channels for model input as shown in Fig. 3. During preprocessing, a region-of-interest (ROI) filter was applied to determine whether each tile should be retained or discarded. All automatically generated ROIs were subsequently subjected to visual inspection to verify annotation accuracy. Tiles exhibiting quality issues—such as tissue folding, edge artifacts, or poor scan resolution—were excluded through manual review of WSI patch location maps. After this curation process, a



Case Control

Fig. 3: Sample images of the dataset used: case group (left) and control group (right).

total of 176,945 high-quality tiles were retained in each class. The final dataset was divided into training (70%), validation (15%), and testing (15%) subsets.

1) Data Management Strategies: Patients without known high-risk clinical factors for CRC, in which low-grade tubular adenomas were identified during screening colonoscopy, were included in the study. A total of 81 patients (41 male, 40 female), ranging in age from 54-95 years (average 70), underwent at least one screening colonoscopy with associated biopsies demonstrating tubular adenomas with low-grade dysplasia; no biopsies showed any histologic features that were indicative of high-risk progression to CRC. Patients were stratified into two cohorts: a precancer group and a control group. The case group consisted of individuals who subsequently developed CRC following screening colonoscopies in which low-grade tubular adenomas were identified. The control group comprised of individuals with no history of CRC despite having low-grade adenomas detected on one or more screening procedures. Compared to the case group, the control group had a greater average number of biopsies and a longer mean screening interval. On average, patients in the case group were 6.86 years older than those in the control group. Histologic slides from both groups containing low-grade tubular adenomas were digitized using the same Leica Aperio AT2 whole slide scanner to generate image data for this study.

E. Results

Quantitative comparisons—Accuracy, F1, Precision and Recall—for colorectal adenoma classification using ViT, Swin Transformer, and UltraLight Med-Vision Mamba are summarized in Table II. While the transformer-based models (ViT and Swin Transformer) primarily utilize self-attention mechanisms to model long-range dependencies across image patches achieved comparable performance with accuracies of 89.84% and 89.52% respectively. In contrast, UltraLight Med-Vision Mamba employs State Space Models (SSMs), which process sequences bidirectionally. This approach allows UltraLight Med-Vision Mamba to achieve 97.34% with higher F1, Precision and Recall, indicating balanced and robust performance. Its ability to better capture subtle dependencies within high-resolution images offered it a distinct advantage. The incorporation of the SCAB module enhanced the feature propagation required for the image classification task. As the SSM models the hidden states over time, it offers the ability to excel in modeling long- and short-range dependencies.

TABLE II: Quantitative performances of ViT, Swin Transformer and UltraLight Med-Vision Mamba.

Model	Accuracy	F1	Precision	Recall
Vision Transformer	89.84%	0.8920	0.9519	0.8392
Swin Transformer	89.52%	0.8878	0.9548	0.8296
UltraLight Med-Vision Mamba	**97.34%**	**0.9733**	**0.9780**	**0.9686**

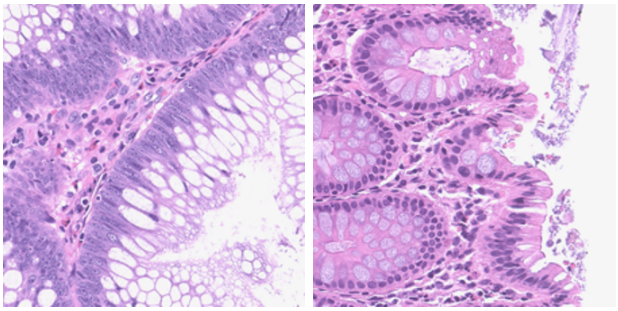
IV. Discussion

This section discusses the limitations of the model's ability to generalize and interpret in colorectal histopathology. The misclassified predictions from each of the three architectures–UltraLight Med-Vision Mamba, ViT, and Swin Transformer–are shown in Figures 4 to 6. This illustrates the visual and histological features that may have contributed to incorrect classifications.

Predicted outputs of UltraLight Med-Vision Mamba: In Fig. 4 (a), the patch clearly demonstrates features of tubular adenoma: nuclear pseudostratification, hyperchromatic elongated nuclei, and goblet cell depletion. Despite the cytological hallmarks being present, the model misclassified this region as class control. This may perhaps reflect the model's limited sensitivity to focal dysplasia, particularly in areas with mixed histology, where adjacent non-dysplastic crypts can mask subtle neoplastic changes. These results may highlight the need for enhanced model training with finer-grained annotations and increased representation of early or borderline dysplasia cases.

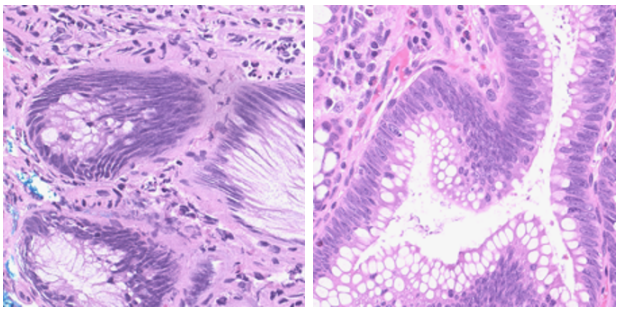
In Fig. 4 (b), the region is histologically consistent with slight adenomatous changes from the control group and was misclassified by the model as a tubular adenoma that progressed to CRC (case group). The crypts are well-formed, with preserved spacing, abundant goblet cells, and basally aligned nuclei, lacking pseudostratification or any cytologic atypia that may be indicative of a high-grade dysplasia, or any other signs of progression to CRC. This false positive may be due to subtle visual cues such as epithelial overcrowding near the tissue edge or darker nuclear staining, which the model may overfit during training.

Predicted outputs of Vision Transformer: The Fig. 5 (a) is an image from the case group that was misclassified with a control group prediction. The model's benign interpretation was likely influenced by confounding factors such as tangential plane of sectioning and the absence of pronounced pseudostratification. The preserved goblet cells and rounded glandular contours may have biased the classifier towards a benign interpretation. This suggests that architectural orientation and crypt profile may significantly influence model sensitivity to dysplasia.



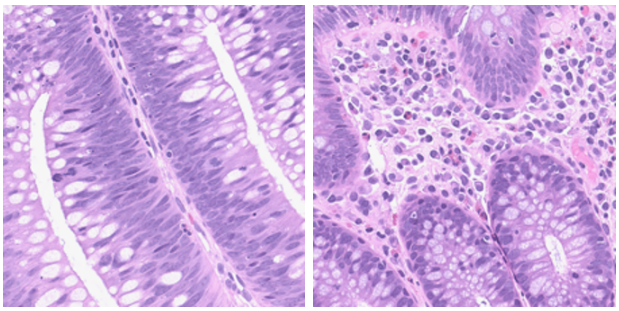
a.) Target: Case, Predicted: Control b.) Target: Control, Predicted: Case

Fig. 4: Mismatched predicted outputs of UltraLight Med-Vision Mamba.



a.) Target: Case, Predicted: Control b.) Target: Control, Predicted: Case

Fig. 5: Mismatched predicted outputs of Vision Transformer.



a.) Target: Case, Predicted: Control b.) Target: Control, Predicted: Case

Fig. 6: Mismatched predicted outputs of Swin Transformer.

Figure 5 (b) was expected to be classified as control, but was misclassified as case by the model. The likely reason lies in the cytologic pseudostratification observed in the glandular epithelium, which mimics low-grade dysplasia. The elongated nuclei aligned perpendicularly to the basement membrane, as well as the increase in nuclear crowding, may have falsely signaled dysplastic changes in the classifier.

Predicted outputs of Swin Transformer:

A false negative occurred for Fig. 6 (a), which represents a tubular adenoma from the case group, perhaps due to its bland architecture and relatively preserved nuclear polarity. While mild nuclear elongation as well as pseudostratification are indeed present, the retention of goblet cells and lack of architectural complexity may have masked dysplastic cues from the classifier. This highlights the challenge of detecting low-grade adenomas that closely mimic normal mucosa at a higher magnification.

A false positive classification was made for Fig. 6 (b), originally belonging in the control group. The classifier likely over-weighted features of mild atypia associated with inflammation due to an influx of lymphocytes and plasma cells in one of the colonic layers. In the real-life practice of pathology, it is common for these reactive epithelial changes to mimic dysplastic characteristics from the case group, leading to erroneous prediction(s). This case highlights a critical challenge for classification of histopathological specimens: discerning true dysplastic (and in some applications, neoplastic) changes from a wide range of confounding inflammatory and reactive processes.

V. Conclusion

This study demonstrates the potential of the UltraLight Med-Vision Mamba architecture for improving the classification of low-grade colorectal adenomas from whole slide images. By effectively modeling long and short-range dependencies and complex spatial relationships through Parallel UltraLight Med-Vision Mamba layers, the network captures subtle histological patterns by enhancing feature extraction that conventional methods may overlook.

Acknowledgment

The authors would like to thank the South Bend Medical Foundation for generously providing the dataset and medical insight required in this study.

References

[1] R. L. Siegel, N. S. Wagle, A. Cercek, R. A. Smith, and A. Jemal, "Colorectal cancer statistics, 2023," *CA: A Cancer Journal for Clinicians*, vol. 73, no. 3, p. 233–254, 2023. [Online]. Available: https://acsjournals.onlinelibrary.wiley.com/doi/abs/10.3322/caac.21772

[2] J. Rosai, *Rosai and Ackerman's Surgical Pathology, 10e*, 10th ed. Elsevier, Jul. 2011, vol. 1.

[3] E. Fearon and B. Vogelstein, "A genetic model for colorectal tumorigenesis," *Cell*, vol. 61, no. 5, pp. 759–767, Jun. 1990.

[4] U. P. S. T. Force, "Screening for colorectal cancer: Us preventive services task force recommendation statement," *JAMA*, vol. 325, no. 19, p. 1965–1977, May 2021. [Online]. Available: https://doi.org/10.1001/jama.2021.6238

[5] J. Lee, C. Jensen, T. Levin, C. Doubeni, A. Zauber, J. Chubak, A. Kamineni, J. Schottinger, N. Ghai, N. Udaltsova, W. Zhao, B. Fireman, C. Quesenberry, E. Orav, C. Skinner, E. Halm, and D. Corley, "Long-term risk of colorectal cancer and related death after adenoma removal in a large, community-based population," *Gastroenterology*, vol. 158, no. 4, pp. 884–894.e5, Mar. 2020, publisher Copyright: © 2020 AGA Institute.

[6] B. Korbar, A. M. Olofson, A. P. Miraflor, K. M. Nicka, M. A. Suriawinata, L. Torresani, A. A. Suriawinata, and S. Hassanpour, "Deep-learning for classification of colorectal polyps on whole-slide images," 2017. [Online]. Available: https://arxiv.org/abs/1703.01550

[7] R. Wu, Y. Liu, P. Liang, and Q. Chang, "Ultralight vm-unet: Parallel vision mamba significantly reduces parameters for skin lesion segmentation," 2024. [Online]. Available: https://arxiv.org/abs/2403.20035

[8] L. Zhu, B. Liao, Q. Zhang, X. Wang, W. Liu, and X. Wang, "Vision mamba: Efficient visual representation learning with bidirectional state space model," 2024. [Online]. Available: https://arxiv.org/abs/2401.09417

[9] A. Sultana, S. N. Abouzahra, V. K. Asari, T. Aspiras, R. Liu, I. Sudakow, and L. Cooper, "Ultralight visionmamba unet: a segmentation architecture for meltpond region localization," in *Pattern Recognition and Prediction XXXVI*, M. S. Alam and V. K. Asari, Eds., vol. 13464. SPIE, 2025, p. 134640N, backup Publisher: International Society for Optics and Photonics. [Online]. Available: https://doi.org/10.1117/12.3054674

[10] J. Ruan, S. Xiang, M. Xie, T. Liu, and Y. Fu, "Malunet: A multi-attention and light-weight unet for skin lesion segmentation," 2022. [Online]. Available: https://arxiv.org/abs/2211.01784

[11] 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," *CoRR*, vol. abs/2010.11929, 2020. [Online]. Available: https://arxiv.org/abs/2010.11929

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2023. [Online]. Available: https://arxiv.org/abs/1706.03762

[13] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," *CoRR*, vol. abs/2103.14030, 2021. [Online]. Available: https://arxiv.org/abs/2103.14030