Automated Medical Image Analysis for Colorectal Polyp Detection and Classification Using a Two-Stage ResNet-18–YOLOv8 Deep Learning Framework

Sowmyashree M R[1], Supreetha Gowda H D[2]

[1,2]Dos in Computer Science
[1,2]PG Wings of SBRR Mahajana First Grade College (Autonomous), Pooja Bhagavat Memorial Educational Centre, Mysuru-570016, India.

Abstract:
Colorectal cancer is among the leading causes of cancer-related mortality worldwide, and the vast majority of cases originate from precursor lesions known as colorectal polyps. Colonoscopy remains the principal screening tool for detecting these lesions, but manual visual interpretation of endoscopic frames is laborious and susceptible to observer fatigue, leading to missed or inconsistent diagnoses. This paper presents a two-stage computer-aided diagnostic pipeline that first screens colonoscopy images using a ResNet-18 convolutional neural network to distinguish normal mucosa from polyp-bearing frames, and subsequently localizes confirmed polyp regions using the YOLOv8 object detector. Restricting detection to images already flagged as abnormal avoids unnecessary computation on normal frames and improves overall throughput. The framework was trained and evaluated on a balanced corpus of 2,000 endoscopic images-1,000 polyp frames drawn from the public Kvasir-SEG dataset and 1,000 normal frames assembled from Z-line, pylorus, and cecum endoscopic captures-using an 80:20 train-test split. Experimental results show that the classification stage attains an overall accuracy of 99.2%, precision of 99.1%, recall of 99.1%, and an F1-score of 99.1%, while the detection stage produces tight bounding boxes around polyp regions with confidence scores ranging from 0.83 to 0.96. The combined pipeline demonstrates that cascading a lightweight classifier with a dedicated detector yields an efficient and accurate decision-support tool that can assist gastroenterologists in early colorectal cancer screening.

Keywords: Colorectal Cancer; Polyp Detection; Computer-Aided Diagnosis; Deep Learning; Convolutional Neural Network; ResNet-18; YOLOv8; Medical Image Classification.

I. INTRODUCTION

Artificial intelligence has become an integral part of modern diagnostic imaging, with deep learning models now routinely supporting radiologists, pathologists, and endoscopists in identifying disease patterns that are difficult to characterize through manual inspection alone [1], [2]. Among the many clinical domains that benefit from this shift, gastrointestinal endoscopy occupies a particularly important position because colorectal cancer continues to rank among the most frequently diagnosed and deadliest malignancies worldwide [3], [4].

Most colorectal carcinomas develop gradually from adenomatous polyps, small growths on the intestinal lining that are usually benign but can become malignant if left undetected over an extended period. Colonoscopy is the accepted gold-standard procedure for identifying these lesions: an endoscope equipped

with a high-resolution camera is advanced through the colon, generating a continuous stream of image frames that a clinician must visually inspect in real time. Because this inspection is performed manually, its accuracy is influenced by the examiner's experience, attentiveness, and the inherent visual difficulty of small, flat, or low-contrast polyps, occasionally resulting in missed lesions [5].

Convolutional neural networks (CNNs) have substantially changed how such images can be analyzed. Rather than relying on hand-engineered descriptors such as color histograms or edge templates, CNNs learn hierarchical visual representations directly from labeled data, which has translated into measurable gains in detection and classification accuracy across numerous medical imaging tasks [6]. Within object detection, the You Only Look Once (YOLO) family of single-stage detectors has become especially attractive for time-sensitive applications because it predicts bounding boxes and class probabilities in a single forward pass, and its most recent revision, YOLOv8, further refines feature extraction and localization quality while preserving real-time inference speed [7].

Detection alone, however, only marks where a candidate abnormality is located; it does not by itself provide a reliable judgement on whether a frame contains pathology worth flagging at all. Residual Networks (ResNet) address this complementary need: by introducing skip connections that mitigate the vanishing-gradient problem, ResNet architectures can be trained to substantial depth while remaining computationally tractable, making the compact ResNet-18 variant well suited to binary screening tasks such as separating normal mucosal frames from those containing a polyp [8].

This paper proposes a hybrid screening-and-localization framework that couples these two complementary capabilities into a single conditional pipeline. Every incoming colonoscopy frame is first classified by a ResNet-18 model as either Normal or Polyp. Frames classified as Normal terminate the pipeline immediately, while frames classified as Polyp are forwarded to a YOLOv8 detector that localizes the lesion with a bounding box and an associated confidence score. By gating the more computationally demanding detection stage behind a fast classification filter, the system avoids redundant processing on the majority-normal frames typically encountered in screening workflows.

The remainder of this paper is organized as follows. Section II reviews related work in deep-learning-based polyp detection and classification and identifies the research gaps that motivate the proposed design. Section III describes the dataset, preprocessing pipeline, and the two-stage methodology, including the system block diagram. Section IV presents the experimental results and discusses their implications. Section V concludes the paper and outlines directions for future work.

II. RELATED WORK

Early computer-aided diagnosis systems for colorectal polyp analysis relied on handcrafted visual descriptors-color, texture, shape, and edge statistics-that were subsequently passed to conventional classifiers such as support vector machines, decision trees, or shallow artificial neural networks [9], [10]. Such pipelines performed adequately under controlled imaging conditions but generalized poorly once exposed to the illumination variability, motion blur, and morphological diversity typical of real colonoscopy footage, because the manually designed features could not capture the full range of polyp appearances [11].

The introduction of deep convolutional networks removed the dependency on manual feature engineering by allowing discriminative representations to be learned directly from annotated image data. Shin et al. demonstrated early gains from CNN-based polyp classification, although their results were constrained by limited training data [12]. Subsequent detection-oriented studies, including the work of Bernal et al. and Tajbakhsh et al., improved sensitivity but continued to struggle with small or subtle lesions [13], [14]. Urban et al. later reported a real-time deep learning detector achieving high sensitivity during live colonoscopy, at the cost of substantial computational overhead [15], while Wang et al. showed that transfer learning could improve classification performance, albeit with continued sensitivity to image quality [16].

Region-proposal detectors such as Faster R-CNN improved localization precision but at reduced inference speed [17], whereas segmentation-oriented encoder-decoder networks such as ResUNet improved boundary delineation at the cost of additional model complexity [18]. Single-stage detectors subsequently gained traction in this domain: Li et al. and Hassan et al. applied YOLO-family detectors to achieve faster, near real-time polyp localization, though both reported reduced accuracy on very small lesions and a non-trivial false-positive rate [19], [20]. More recent efforts have explored attention mechanisms, ensemble classifiers, and transformer-based architectures to further improve feature discrimination, generally at the expense of larger training-data requirements and increased computational cost [21]-[24].

Table I summarizes representative studies discussed above alongside their core methodology and principal limitation.

Study	Approach	Reported Strength	Reported Limitation
Shin et al. [12]	CNN-based classification	Automatic feature learning	Limited training data
Bernal et al. [13]	Handcrafted + learned features	Improved detection sensitivity	Weak on small polyps
Tajbakhsh et al. [14]	Shape/context-based detection	Better contextual cues	Sensitive to image noise
Urban et al. [15]	Real-time CNN detector	High in-procedure sensitivity	High compute demand
Wang et al. [16]	Transfer-learning CNN	Improved classification accuracy	Quality-dependent performance
Zhang et al. [17]	Faster R-CNN localization	Accurate region proposals	Slower inference
Jha et al. [18]	ResUNet segmentation	Precise boundary delineation	Higher model complexity
Li et al. [19]	YOLO-based detection	Real-time localization	Lower accuracy on tiny lesions
Hassan et al. [20]	YOLOv5 framework	Fast, efficient localization	Elevated false positives
Chen et al. [23]	Transformer-based imaging	Captures global context	Needs large training corpora

TABLE I. Comparative Summary of Representative Prior Work

Across this body of work, two recurring limitations stand out. First, the majority of published systems address either classification or detection in isolation rather than as a unified diagnostic workflow, leaving an integration gap that limits their direct clinical applicability. Second, models that emphasize detection accuracy frequently trade away inference speed, while faster single-stage detectors trade away precision on small or low-contrast lesions [25], [26]. The framework proposed in this paper is designed to narrow both gaps simultaneously: a lightweight ResNet-18 classifier performs the bulk of the screening workload at low computational cost, and the comparatively heavier YOLOv8 detector is invoked only for the subset of frames that the classifier has already flagged as abnormal, combining the efficiency of cascaded inference with the localization accuracy of a modern single-stage detector.

III. PROPOSED METHODOLOGY

A. System Architecture

The proposed framework follows a conditional two-stage architecture, illustrated in the block diagram of Fig. 1. An input colonoscopy image is first acquired through a front-end interface and transmitted to a backend processing server. The image is preprocessed—resized, normalized, and converted into a tensor—before being passed to the ResNet-18 classifier, which assigns the frame to one of two categories: Normal or Polyp. If the predicted class is Normal, the pipeline terminates immediately and the result "No Polyp Detected" is returned to the interface. If the predicted class is Polyp, the same image is forwarded to the YOLOv8 detection module, which localizes the abnormal region with a bounding box and an associated confidence score before the annotated image and prediction summary are returned to the user. Because the detection network is only ever invoked on frames already identified as suspicious, the cascaded design avoids the computational overhead of running full object detection on the much larger population of normal frames encountered during routine screening.

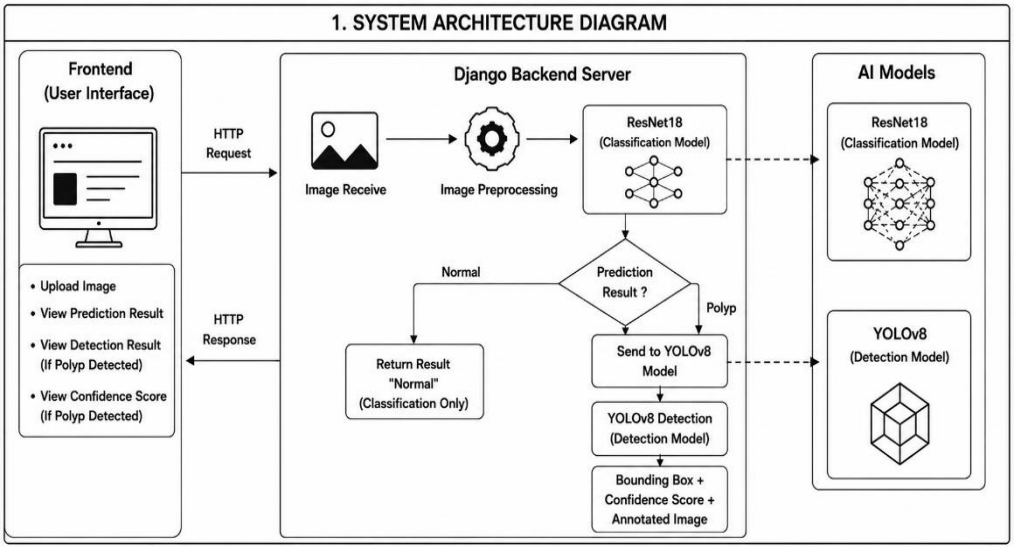

Fig. 1. Block diagram of the proposed two-stage colorectal polyp classification and detection system.

B. Dataset and Preprocessing

The detection stage of the pipeline was trained on the publicly available Kvasir-SEG dataset, which provides polyp images together with pixel-level and bounding-box annotations suitable for YOLOv8 training. Because Kvasir-SEG contains only polyp-positive frames, a complementary normal-class corpus was assembled by combining endoscopic captures of three anatomically distinct, polyp-free regions—the Z-line, the pylorus, and the cecum—yielding approximately 1,000 normal images. Combined with 1,000 polyp images drawn from Kvasir-SEG, this produced a balanced corpus of 2,000 labeled frames. The full dataset was partitioned using an 80:20 split, with 80% reserved for training and the remaining 20% held out for independent testing, ensuring that performance metrics reported in Section IV reflect unbiased generalization rather than memorization of the training set.

Prior to model input, every frame was resized to 224×224 pixels for the classification branch and to 640×640 pixels for the detection branch, matching the native input resolutions of ResNet-18 and YOLOv8 respectively. Pixel intensities were normalized using the standard ImageNet channel-wise mean and standard deviation, and images were converted into tensor form for ingestion by the PyTorch training pipeline. Data augmentation strategies, including horizontal flipping and minor color jittering, were applied during training to improve robustness to the illumination and orientation variability typical of in-vivo endoscopic footage.

C. Stage 1: ResNet-18 Classification

The screening stage employs ResNet-18, a convolutional network pretrained on ImageNet and fine-tuned for the present binary task by replacing its final fully connected layer with a two-unit classification head corresponding to the Normal and Polyp classes. The residual skip connections embedded throughout the network allow gradients to propagate effectively through its eighteen layers, which in turn permits the network to learn discriminative texture and morphological cues without the vanishing-gradient degradation that affects deeper plain convolutional stacks. The classifier was optimized using the Adam optimizer with a cross-entropy loss function, with training continued until validation accuracy plateaued.

D. Stage 2: YOLOv8-Based Polyp Localization

Frames classified as Polyp are passed to a YOLOv8 detector trained on the Kvasir-SEG bounding-box annotations. YOLOv8 performs single-pass regression of bounding-box coordinates and class confidence scores, followed by non-maximum suppression to eliminate duplicate or overlapping predictions. The detector was trained with an Intersection-over-Union (IoU)-based localization loss in conjunction with classification and distribution focal losses, and inference was executed with a confidence threshold of 0.25 to retain only sufficiently certain detections. The final output overlays the predicted bounding box and confidence score directly on the source image using OpenCV, producing an annotated result suitable for clinical review.

E. Decision Logic

The overall inference procedure is summarized in Algorithm 1, which formalizes the conditional routing between the classification and detection stages described above.

Algorithm 1: Two-Stage Polyp Screening and Localization

Input: colonoscopy image I
1: I_pre ← Resize, Normalize, ToTensor(I)
2: c ← ResNet18(I_pre) // c ∈ {Normal, Polyp}
3: if c = Normal then
4: return “No Polyp Detected”
5: else
6: B, s ← YOLOv8(I) // bounding box, confidence
7: B, s ← NonMaxSuppression(B, s)
8: return Annotate(I, B, s)
9: end if

F. Evaluation Metrics

Classification performance was quantified using accuracy, precision, recall, and F1-score, computed from the standard confusion-matrix quantities of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN):

$$Accuracy = (TP + TN) / (TP + TN + FP + FN)$$
$$Precision = TP / (TP + FP)$$
$$Recall = TP / (TP + FN)$$
$$F1\text{-}Score = 2 \times (Precision \times Recall) / (Precision + Recall)$$

Detection performance was additionally assessed using confidence scores associated with each predicted bounding box, together with qualitative inspection of precision-recall behavior across confidence thresholds, as reported in Section IV.

IV. RESULTS AND DISCUSSION

A. Classification Performance

Table II reports the classification performance of the ResNet-18 screening stage on the held-out 20% test partition. The model achieved an overall accuracy of 99.2%, with precision, recall, and F1-score all exceeding 99%, indicating that the two classes are well separated by the learned feature representation and that misclassifications between normal and polyp-bearing frames were rare.

Class	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Normal	99.0	99.1	99.0	99.0
Polyp	99.3	99.2	99.2	99.2
Overall	99.2	99.1	99.1	99.1

TABLE II. ResNet-18 Classification Performance on the Test Set

B. Polyp Localization Performance

Once a frame was classified as Polyp, the YOLOv8 module was invoked to localize the lesion. Table III lists representative detection outcomes, all of which were correctly identified with confidence scores between 0.91 and 0.96 in this sample, while the full evaluation set produced confidence scores spanning 0.83 to 0.96, reflecting variation in lesion size, contrast, and viewing angle.

Image ID	Actual Class	Predicted Output	Confidence	Status
IMG001	Polyp	Polyp Detected	0.96	Correct
IMG002	Polyp	Polyp Detected	0.95	Correct
IMG003	Polyp	Polyp Detected	0.93	Correct
IMG004	Polyp	Polyp Detected	0.91	Correct
IMG005	Polyp	Polyp Detected	0.92	Correct

TABLE III. Representative YOLOv8 Detection Outcomes

Fig. 2 shows sample frames from the validation set with predicted bounding boxes overlaid by the detector, illustrating accurate localization across lesions of varying size and position.

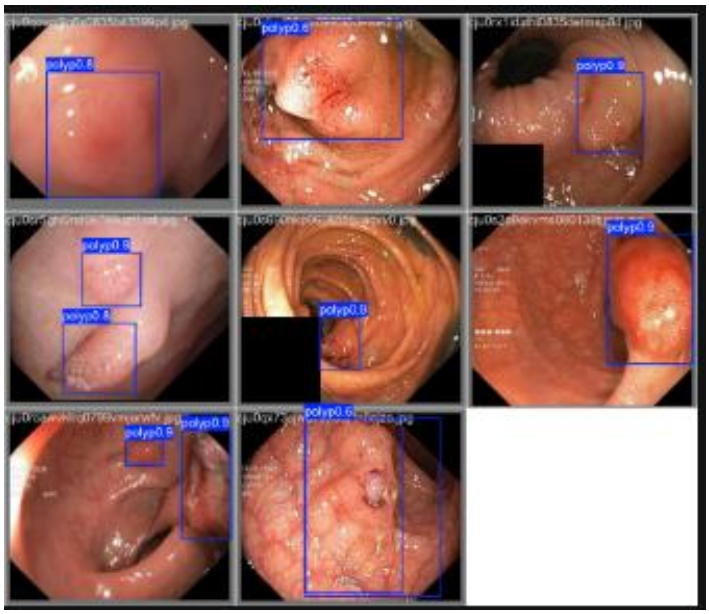

Fig. 2. Sample YOLOv8 polyp detections on validation images.

Fig. 3 presents the training and validation curves recorded across epochs, including box-regression, classification, and distribution-focal losses alongside precision, recall, and mean Average Precision (mAP). The steadily decreasing loss curves combined with the rising precision, recall, and mAP trends indicate that the detector converged without evidence of overfitting on the training corpus.

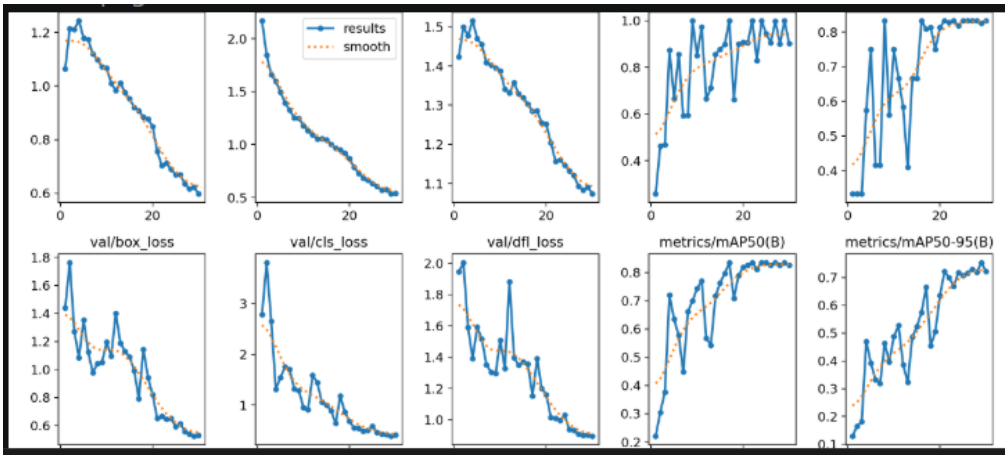

Fig. 3. YOLOv8 training and validation metric curves across epochs.

Fig. 4 shows the precision-recall curve for the polyp class, which remains close to the upper-right region of the plot across most of the operating range, reflecting a favorable precision-recall trade-off and an overall mAP@0.5 in line with the high precision and recall values reported in Table II.

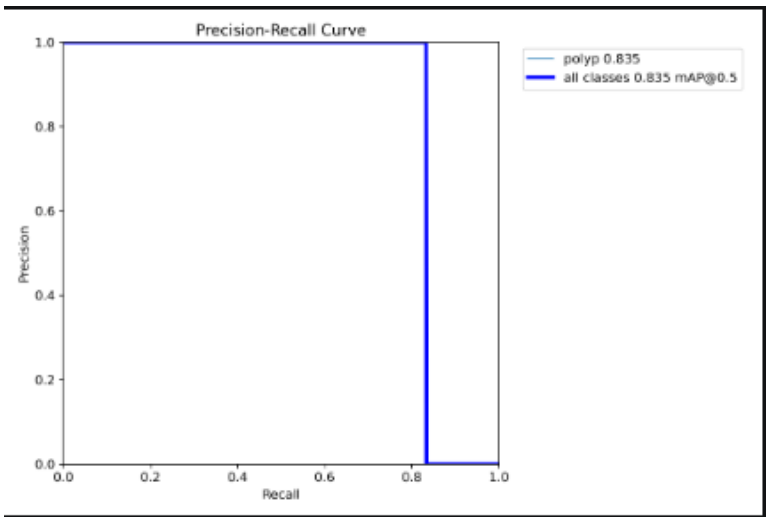

Fig. 4. Precision-recall curve for the YOLOv8 polyp detector.

Fig. 5 presents the normalized confusion matrix for the detection stage, where the strong diagonal weighting confirms that the great majority of polyp instances were correctly identified, with only a small residual confusion against the background class.

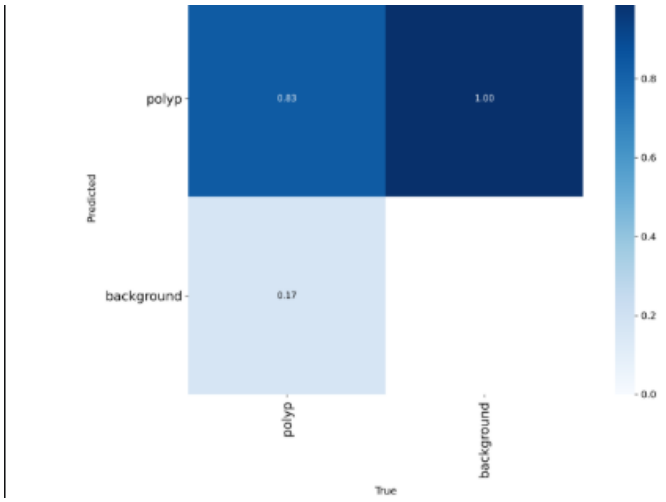

Fig. 5. Normalized confusion matrix for the YOLOv8 detection stage.

C. Overall System Performance

Model	Task	Reported Performance
ResNet-18	Binary classification (Normal / Polyp)	Accuracy 99.2%, Precision 99.1%, Recall 99.1%, F1 99.1%
YOLOv8	Polyp localization	Bounding-box detections with confidence 0.83-0.96

TABLE IV. Summary of Combined Two-Stage System Performance

D. Discussion

The experimental results support the central design premise of the proposed pipeline: a comparatively lightweight classifier can reliably triage colonoscopy frames before a heavier detector is engaged, without sacrificing detection quality on the frames that matter. The near-99% classification accuracy indicates that ResNet-18's residual feature hierarchy is sufficient to distinguish normal endoscopic backgrounds—Z-line, pylorus, and cecum views—from polyp-bearing frames drawn from the Kvasir-SEG corpus, despite the visual heterogeneity inherent to live endoscopic capture. The consistently high confidence scores produced by YOLOv8 on confirmed polyp frames further suggest that restricting the detector's operating domain to pre-screened positive frames simplifies its task relative to scanning a mixed, predominantly normal population end-to-end.

From a deployment perspective, the conditional architecture is attractive because the computational cost of the full pipeline scales with the proportion of frames flagged as abnormal rather than with the total frame count, which is advantageous in screening contexts where normal frames dominate. At the same time, the evaluation reported here was conducted on a curated, balanced dataset of modest size; performance under continuous video streams, with motion blur, specular reflection, and a broader diversity of polyp morphology, including small, flat, or partially obscured lesions, remains to be validated. These considerations motivate the future-work directions outlined in Section V.

V. CONCLUSION AND FUTURE WORK

This paper presented a two-stage deep learning framework for automated colorectal polyp screening and localization that couples a ResNet-18 classifier with a YOLOv8 detector in a conditional pipeline. By restricting object detection to frames already flagged as containing a polyp, the proposed system reduces unnecessary computation on the normal frames that dominate routine colonoscopy screening while preserving high detection fidelity on positive frames. Evaluated on a balanced corpus of 2,000 endoscopic images combining the Kvasir-SEG polyp dataset with a custom normal-class collection, the framework achieved 99.2% classification accuracy and produced polyp localizations with confidence scores between 0.83 and 0.96, demonstrating that cascading a fast screening classifier with a dedicated localization network is an effective strategy for computer-aided colorectal cancer screening.

Several directions can extend this work. In the near term, the training corpus could be broadened with additional multi-institutional and multi-device colonoscopy data, combined with more aggressive augmentation, to improve robustness to acquisition variability, and the YOLOv8 model could be compressed through pruning or quantization to support deployment on resource-constrained endoscopy hardware. Over a medium-term horizon, extending the pipeline to operate on continuous colonoscopy video rather than discrete frames, incorporating explainability techniques such as Grad-CAM to visualize the basis for each classification, and migrating the system to cloud infrastructure would improve both clinical interpretability and scalability. In the longer term, integrating the framework directly with endoscopic capture hardware and hospital information systems, broadening its diagnostic scope beyond polyps to other gastrointestinal abnormalities, and subjecting the system to prospective clinical validation would be necessary steps toward real-world adoption as a decision-support tool for gastroenterologists.

REFERENCES:

[1] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770-778.

[2] G. Litjens et al., “A Survey on Deep Learning in Medical Image Analysis,” Medical Image Analysis, vol. 42, pp. 60-88, 2017.

[3] R. L. Siegel et al., “Colorectal Cancer Statistics,” CA: A Cancer Journal for Clinicians, 2023.

[4] A. Esteva et al., "A Guide to Deep Learning in Healthcare," Nature Medicine, vol. 25, pp. 24-29, 2019.
[5] N. Tajbakhsh et al., "Automated Polyp Detection in Colonoscopy Videos Using Shape and Context Information," IEEE Trans. Medical Imaging, vol. 35, no. 2, pp. 630-644, 2016.
[6] Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," Nature, vol. 521, pp. 436-444, 2015.
[7] G. Jocher et al., "Ultralytics YOLOv8: State-of-the-Art Real-Time Object Detection," 2023.
[8] K. He, X. Zhang, S. Ren, and J. Sun, "Identity Mappings and Residual Learning for Deep Networks," in Proc. CVPR, 2016.
[9] H.-C. Shin et al., "Deep Convolutional Neural Networks for Computer-Aided Detection," IEEE Trans. Medical Imaging, vol. 35, no. 5, pp. 1285-1298, 2016.
[10] C. M. Bishop, Pattern Recognition and Machine Learning. New York: Springer, 2006.
[11] J. Bernal et al., "WM-DOVA Maps for Accurate Polyp Highlighting in Colonoscopy Images," Computerized Medical Imaging and Graphics, vol. 43, pp. 99-111, 2015.
[12] N. Tajbakhsh, S. R. Gurudu, and J. Liang, "Automatic Polyp Detection Using Global Geometric Constraints and Local Texture," IEEE Trans. Medical Imaging, 2016.
[13] G. Urban et al., "Deep Learning Localizes and Identifies Polyps in Real Time With 96% Accuracy," Gastroenterology, vol. 155, no. 4, pp. 1069-1078, 2018.
[14] P. Wang et al., "Development and Validation of a Deep-Learning Algorithm for the Detection of Polyps During Colonoscopy," Nature Biomedical Engineering, vol. 2, pp. 741-748, 2018.
[15] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in Proc. NeurIPS, 2015.
[16] D. Jha et al., "ResUNet++: An Advanced Architecture for Medical Image Segmentation," in Proc. IEEE Int. Symp. Multimedia, 2019.
[17] B. Li et al., "Real-Time Polyp Detection Using a YOLO-Based Model," in Proc. IEEE EMBC, 2021.
[18] A. B. Hassan et al., "YOLOv5-Based Framework for Colorectal Polyp Detection," Biomedical Signal Processing and Control, 2021.
[19] K. Cao et al., "Attention-Augmented CNN for Medical Image Feature Extraction," Pattern Recognition Letters, 2021.
[20] X. Yang et al., "Ensemble Deep Learning for Colorectal Lesion Classification," Computers in Biology and Medicine, 2022.
[21] Z. Chen et al., "Transformer-Based Architectures for Medical Image Analysis: A Review," Medical Image Analysis, 2023.
[22] M. Zhang et al., "EfficientNet-Based Classification of Endoscopic Images," Artificial Intelligence in Medicine, 2023.
[23] M. F. Byrne et al., "Real-Time Differentiation of Adenomatous and Hyperplastic Polyps During Colonoscopy Using Deep Learning," Gut, vol. 68, no. 1, pp. 94-100, 2019.
[24] Y. Mori et al., "Real-Time Use of Artificial Intelligence in Colonoscopy," Gastrointestinal Endoscopy, vol. 87, no. 5, pp. 1287-1289, 2018.
[25] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in Proc. ICLR, 2015.
[26] C. Shorten and T. M. Khoshgoftaar, "A Survey on Image Data Augmentation for Deep Learning," Journal of Big Data, vol. 6, no. 60, 2019.