Expert Systems With Applications 295 (2026) 128756

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

ColonOOD: A complete pipeline for optical diagnosis of colorectal polyps integrating out-of-distribution detection and uncertainty quantification

Sehyun Park [a,1], Dongheon Lee [b,1], Ji Young Lee [c], Jaeyoung Chun [d], Ji Young Chang [e], Eunsu Baek [a], Eun Hyo Jin [f], Hyung-Sin Kim [a,*]

[a] Graduate School of Data Science, Seoul National University, Gwanakro 1, Seoul, 08826, South Korea
[b] Department of Radiology, College of Medicine, Seoul National University, Seoul, South Korea
[c] Health Screening and Promotion Center, College of Medicine, Asan Medical Center, University of Ulsan, Seoul, South Korea
[d] Department of Internal Medicine, Gangnam Severance Hospital, Yonsei University College of Medicine, Seoul, South Korea
[e] Department of Health Promotion Medicine, Ewha Womans University Seoul Hospital, Seoul, South Korea
[f] Department of Internal Medicine, Healthcare Research Institute, Seoul National University Hospital Healthcare System Gangnam Center, Seoul, South Korea

ARTICLE INFO

Keywords:
Colonoscopy
Computer-aided diagnosis (CAD)
Colorectal polyp classification
Out-of-distribution detection
Uncertainty quantification

ABSTRACT

The rising prevalence of colorectal cancer necessitates early and accurate optical diagnosis of colorectal polyps. Despite advances in Computer-Aided Diagnosis (CAD) systems, challenges like data variability and inconsistent clinical performance hinder their widespread use. To address these limitations, we propose ColonOOD, an integrated CAD system for polyp localization, uncertainty-aware polyp classification, and Out-of-Distribution (OOD) polyp detection during colonoscopy. ColonOOD ensures robust classification of adenomatous, hyperplastic, and OOD polyps while providing calibrated uncertainty scores to support clinical decisions. Extensive evaluations across four medical centers and two public datasets demonstrate ColonOOD's strong performance, achieving up to 79.69 % classification and 75.53 % OOD detection accuracy. This system offers reliable insights for endoscopists, marking a significant step toward broader clinical adoption of automated diagnostic tools in colorectal cancer care.

1. Introduction

Colorectal cancer is a leading cause of cancer-related deaths globally, yet fast and accurate detection of polyps during colonoscopy has been shown to reduce mortality rates by up to 53 % (Kiwan et al., 2022).

Characterizing these polyps remains challenging and often relies heavily on the expertise of endoscopists. However, factors like inter-expert variability in diagnoses (Jin et al., 2020), the extensive training required for proficiency (Seo et al., 2020), and the high prevalence of minor types of polyps in Fig. 1 beyond the two main types–adenomatous polyps (AD) and hyperplastic polyps (HP)–contribute to significant disparities in diagnostic accuracy and quality of care (Patel et al., 2020).

While histopathological assays are accurate tools for diagnosis, the high costs associated with evaluating polyps that often do not require further screening, such as hyperplastic polyps (HP), have prompted the American Society for Gastrointestinal Endoscopy (ASGE) to recommend the optical classification of diminutive polyps during colonoscopy and their subsequent discard without histopathological analysis (Parsa et al., 2021). This strategy can save up to $1.06 billion annually (Patel et al., 2020), while also enabling immediate clinical decision-making regarding treatment and patient management.

These challenges and need highlight the pressing need for reliable diagnostic tools during colonoscopy, such as computer-aided diagnosis (CAD) systems, to enhance the optical diagnosis of colorectal polyps by enabling rapid and accurate decision-making.

Despite their potential, CAD systems face substantial challenges in real-world deployment. Existing systems often lack consideration for clinical complexities during optical diagnostics, such as unseen polyp types and inconsistencies between CAD and clinician decisions (Beger et al., 2021; Chen et al., 2021; Wang et al., 2020). Research has focused primarily on the binary classification of AD and HP polyps (Lo et al., 2022; Yoshida et al., 2021), or it has expanded to a few limited additional categories such as advanced adenomas and serrated polyps (Ozawa et al., 2020; Yang et al., 2020). However, these efforts frequently fail to address the full spectrum of polyp diversity (Urban et al., 2018) that might occur in the practical workflows of clinical use

* Corresponding author.
E-mail addresses: heon843@snu.ac.kr (S. Park), dhlee13@snu.ac.kr (D. Lee), jiyounglee@amc.seoul.kr (J.Y. Lee), chunjmd@yuhs.ac (J. Chun), 01022s@eumc.ac.kr (J.Y. Chang), beshu9407@snu.ac.kr (E. Baek), icetea@snuh.org (E.H. Jin), hyungkim@snu.ac.kr (H. Kim).
[1] These authors have contributed equally to this work.

https://doi.org/10.1016/j.eswa.2025.128756
Received 2 February 2025; Received in revised form 2 June 2025; Accepted 23 June 2025
Available online 29 June 2025

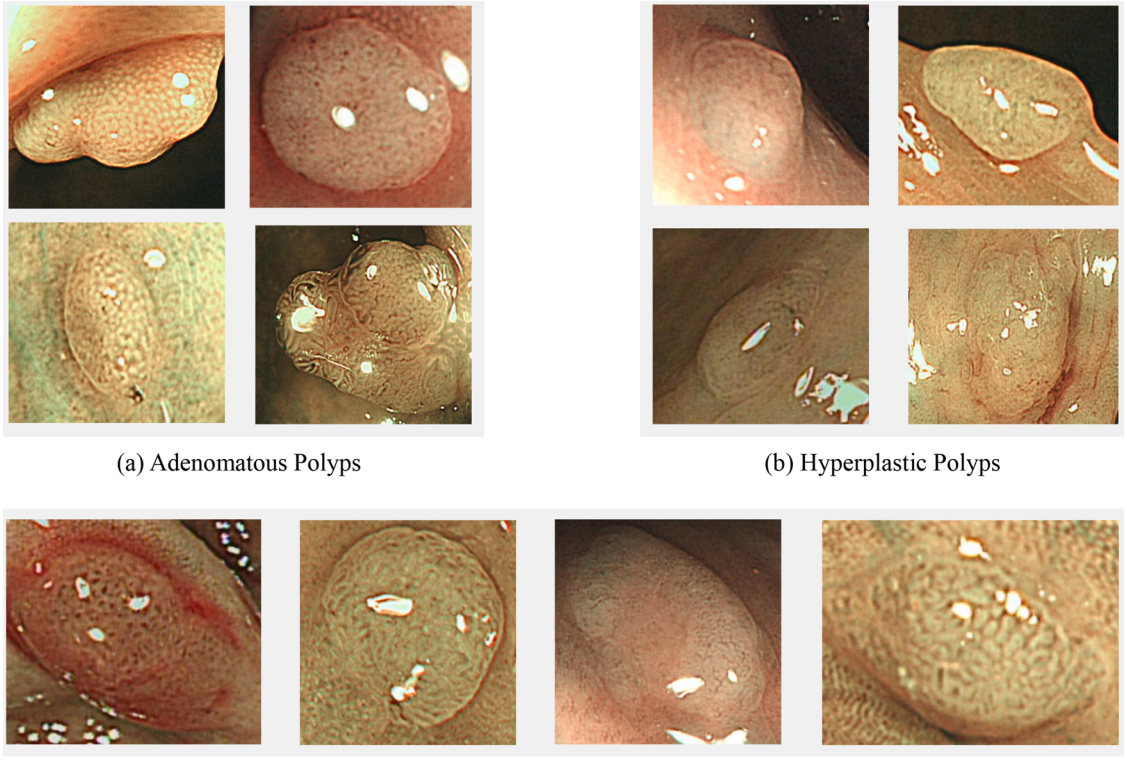
(Wang et al., 2020). These gaps can result in silent failures, undermining prediction reliability and posing critical barriers to widespread adoption (Hong et al., 2024).

A particularly pressing issue is the detection of Out-of-Distribution (OOD) polyps, which include rare or newly emerging types that deviate from known patterns (Jin et al., 2020). These atypical polyps present challenges for both endoscopists and CAD systems, increasing the risk of diagnostic errors and inappropriate treatments (Smith et al., 2024). For example, a malignant OOD polyp misclassified as a benign hyperplastic polyp could delay necessary interventions, potentially jeopardizing patient outcomes. Effective detection of OOD polyps is thus crucial for improving the robustness and clinical reliability of CAD systems in colonoscopy.

This study presents ColonOOD, a comprehensive pipeline addressing key challenges, particularly unseen polyp types, in deploying CAD systems for colonoscopy. The proposed framework integrates polyp localization, classification, and OOD detection to provide a solution tailored for real-time colonoscopy applications. ColonOOD leverages a highly discriminative classification model trained with advanced techniques and an OOD detection module developed through rigorous analysis of user scenarios. By identifying deviations in AD, HP, and OOD polyps and quantifying predictive uncertainty, ColonOOD provides clinically meaningful interpretability for helping to mitigate potential discrepancies between endoscopists' judgments and CAD predictions. These capabilities support endoscopists in making informed decisions, reducing reliance on subjective judgment, and addressing practical challenges in clinical workflows such as unseen polyps.

The specific objectives of this study are threefold: (1) Develop an integrated pipeline for polyp localization, classification, and OOD detection to provide complete pipeline of CAD in real-time colonoscopy; (2) Create an uncertainty-aware classification model capable of distinguishing AD and HP polyps with high accuracy and well-calibrated uncertainty predictions for providing metric to compare endoscopist's decision and CAD's decision; and (3) Design a scenario-targeted OOD detection method to identify and manage rare and unseen polyps for accommodating diverse clinical challenges and enhancing applicability in practice.

- **A complete CAD system for optical diagnosis of colorectal polyps (ColonOOD):** This research introduces a systematic design named ColonOOD, comprised of a robust classification model and scenario-targeted OOD module. ColonOOD systematically localizes polyps, classifies them with uncertainty scores, categorizes predictions into high- or low-confidence groups, and detects OOD polyps. By addressing covariate shifts, inconsistent performance across institutions, and real-world applicability, ColonOOD provides a practical approach for clinical deployment to ensure robust performance in diverse clinical settings and to aid endoscopists in making informed decisions.
- **Out-of-distribution (OOD) polyp detection:** We propose a scenario-targeted OOD detection framework, informed by an in-depth analysis of baseline classification model behaviors and leveraging feature vectors, logit tensors, and probability distributions of deep learning models. To our knowledge, this is the first CAD system for colorectal polyps incorporating OOD detection. It addresses the challenge of detecting rare, newly emerging, or unseen polyp types, thereby improving generalizability across diverse clinical scenarios.
- **Uncertainty-aware polyp classification:** We designed an uncertainty-aware systematic pipeline with a high-accuracy classification model and well-calibrated confidence scores by employing extensive data augmentation, diverse training strategies, optimal model architecture selection, temperature scaling, and uncertainty thresholding. These techniques mitigate overconfidence, enabling endoscopists to interpret AI predictions alongside their clinical judgment, thereby reducing the likelihood of misdiagnosis and enhancing trust in the system.
- **Evaluation considering diverse clinical environments:** To rigorously evaluate ColonOOD, we curated diverse evaluation datasets



(a) Adenomatous Polyps

(b) Hyperplastic Polyps

(c) Other Polyp Types – Traditional Serrated Adenoma, Sessile Serrated Polyps, Serrated Polyps and Others

Fig. 1. Representative image of colorectal polyps. (a) Adenomatous polyps (b) Hyperplastic polyps (c) Other types of polyps such as traditional serrated adenoma, sessile serrated polyps, serrated polyps, and more.

Table 1
Number of adenomas or hyperplastic polyp images from each medical institution.

	Center A (Train)	Center A (Val)	Center B	Center C	Center D
Adenomatous Polyps	1100	278	242	105	105
Hyperplastic Polyps	1050	94	58	73	39

from four medical institutions and two publicly available datasets to give an adequate evaluation of the system's feasibility in diverse clinical settings. Additionally, we developed a demo for simulating and assessing the entire pipeline in colonoscopy videos, replicating real-world clinical scenarios to validate the system's applicability.

2. Materials and methods

2.1. Dataset

Polyp classification dataset Dataset used in our experiment, summarized in Table 1, was collected through four medical institutions: Center A, Center B, Center C and Center D. All data was collected from colonoscopies using Narrow Band Imaging (NBI) and preprocessed by cropping polyps from colonoscopy video frames. As a result, all the input data for the model consisted of polyp ROI images from colonoscopy views. Only data from Center A were used for training and validation, while datasets from other institutions were reserved for external validation.

Uncertainty quantification polyp dataset Ten expert endoscopists evaluated 720 polyp images from four institutions, categorizing the ground truth labels with High Confidence (HC) and Low Confidence (LC). If all ten experts diagnosed a polyp with high confidence, it received a confidence score of 10 HCs. This dataset evaluates the uncertainty and confidence scores of the model and assesses its feasibility for clinical application.

Out-of-distribution polyp dataset The OOD polyp dataset shown in Table 2 includes a diverse range of polyp types, such as *Sessile Serrated Polyps*, *Traditional Serrated Adenomatous*, *Serrated Polyps*, and others. The datasets consist of private collections from Center A and Center A (Video). At the same time, the SUN (Itoh et al., 2020) and POLAR (POL, 2020) datasets are publicly available and used for polyp classification and localization. The OOD polyps were manually extracted from colonoscopy videos using the labels provided in the Center A (Video) and SUN datasets.

Colonoscopy video Center A collected 200 colonoscopy videos, each annotated with the time of polyp emergence and the type of polyps present. This dataset evaluated the model in a clinical setting, closely reflecting actual colonoscopy procedures.

Table 2
Number of Out-of-Distribution Polyps in each dataset. (TSA: Traditional Serrated Adenoma, SSL: Sessile Serrated Polyps, SP: Serrated Polyps, Others include Invasive cancer and others specified by POLAR dataset).

	Internal validation data		External validation data (Public Data)	
	Center A	Center A (Video)	SUN	POLAR
TSA	3	1	2	5
SP	64	10	–	–
SSL	19	2	4	199
Normal Mucosa	–	–	–	26
Other	–	–	1	15
Total	86	13	7	245

2.2. User scenario

Designing a system for OOD detection is highly dependent on the specific application. Firstly, a deep understanding of the context of the problem should be preceded to make a robust and appropriate method for OOD selection. In our scenario depicted in Fig. 2, among all diagnosis cases, an accurate diagnosis of HP is essential, as many OOD polyps are precancerous and therefore require histopathological tests for post-treatment. We may discard HP, while OOD and AD must proceed to post-process. Thus, accurately distinguishing HP from OOD polyps and accurately classifying AD should be the model's priority.

3. Proposed framework: ColonOOD

3.1. Integrated framework

Considering the scenario described above, this study proposes a robust computer-aided diagnosis (CAD) system named ColonOOD, as shown in Fig. 3. This pipeline is structured around three core components: Polyp localization, uncertainty-aware polyp classification, and Out-of-Distribution (OOD) polyp detection.

Once the polyp localization model predicts a cropped image of the polyp, the baseline classification model classifies the polyp as either AD or HP and quantifies the uncertainty of its predictions, producing preliminary confidence scores. If the model classifies a polyp as AD with a high confidence score, the result is forwarded directly to the user interface for immediate display. For cases where polyps are classified as AD with a low confidence score or identified as OOD, the OOD detection module re-evaluates the input to ensure robust decision-making. The OOD detection module consists of an OOD-specialized model and a feature-based post-processing step, which complement each other to effectively distinguish between HP and OOD polyps in a collectively exhaustive manner. By integrating classification and OOD detection module, ColonOOD optimizes boundary decisions and enhances clinical applicability. The framework is summarized in Fig. 3.

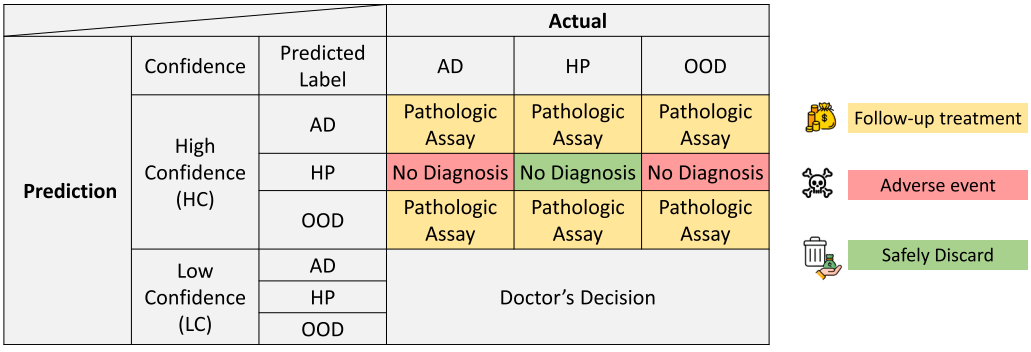


Fig. 2. A possible scenario in real clinical settings.

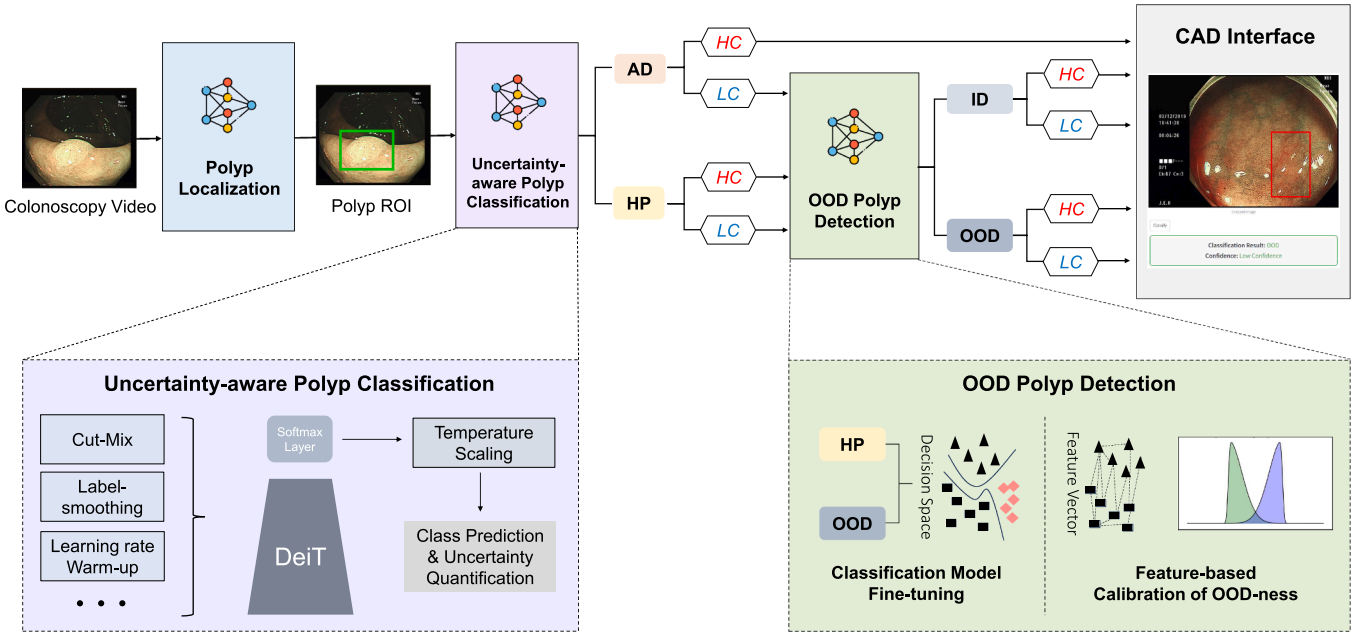


Fig. 3. The overview of the proposed framework named ColonOOD.

3.2. Decision rule

Let $\mathbf{I}$ represent the input-cropped image of the polyp. The baseline classification model generates a predicted label $\hat{y} \in \{\text{AD}, \text{HP}\}$ and an associated confidence score $S(\mathbf{I}) \in [0, 1]$. This process is formalized as:

$$C(\mathbf{I}) = \{\hat{y}, S(\mathbf{I})\}, \tag{1}$$

where:

$$\hat{y} = \arg \max_{c \in \{\text{AD},\text{HP}\}} P(c \mid \mathbf{I}),$$

$$S(\mathbf{I}) = P(\hat{y} \mid \mathbf{I}).$$

The decision-making process integrates classification and OOD detection, defined as:

$$f(\mathbf{I}) = \begin{cases} \{\hat{y}, \text{High Confidence}\}, & \text{if } S(\mathbf{I}) \geq T \text{ and } D(\phi(\mathbf{I})) \geq T_{\text{OOD}}, \\ \{\text{OOD}\}, & \text{if } D(\phi(\mathbf{I})) < T_{\text{OOD}}, \\ \{\hat{y}, \text{Low Confidence}\}, & \text{if } S(\mathbf{I}) < T \text{ and } D(\phi(\mathbf{I})) \geq T_{\text{OOD}}. \end{cases} \tag{2}$$

Here, T is the confidence threshold for classification. $\phi(\mathbf{I})$ maps the input image to a latent feature space. $D(\phi(\mathbf{I}))$ measures the OOD score based on the feature embedding of the input sample. A higher OOD score means that the sample will likely be an ID sample, following the convention of OOD studies (Hendrycks & Gimpel, 2017) (Liang et al., 2018) (Hendrycks et al., 2022). T_{OOD} is the threshold for distinguishing OOD samples.

3.3. Polyp localization

The system is provided with a frame from a colonoscopy in a clinical setting, as shown on the left side of Fig. 3. Since the classification model requires images with localized polyps, we offer two options: (1) clinicians can manually select the region of interest by clicking the top-left and bottom-right corners of the polyp, or (2) the system can automatically localize the polyp using a deep learning model. For the manual localization, we implemented an OpenCV (v3.1) as a tool, allowing the user to click the ROI and crop it into a bounding box image. For automatic localization, we employed pre-trained YOLO-OB, a specialized architecture that features bidirectional multiscale feature fusion and anchor-free box regression, ensuring robust polyp localization (Yang et al., 2023).

3.4. Uncertainty-aware polyp classification

3.4.1. Model and training strategy

The classification model training process incorporates several strategies to mitigate the over-confidence issue caused by the limited amount of data. First, the classification baseline model employed a transformer-based model instead of traditional CNN-based models. Although CNNs have been widely used in the medical field due to their ability to exploit spatial location and their relatively low data requirements (Chan & Siegel, 2018), transformers have consistently shown superior performance in addressing overconfidence and generalizing between different medical institutions (Raghu et al., 2021). Vision transformers (ViTs), in particular, achieve higher accuracy and robustness, making them especially wellsuited for classifying colorectal polyps.

Secondly, the classification model of ColonOOD applied advanced image augmentation techniques such as Mixup (Zhang et al., 2018) to compensate for the limited data and reduce overfitting. Given two examples (x_i, y_i) and (x_j, y_j), MixUp generates a new training example $(\tilde{x}, \tilde{y})$ as follows:

$$\tilde{x} = \lambda x_i + (1 - \lambda) x_j \tag{3}$$

$$\tilde{y} = \lambda y_i + (1 - \lambda) y_j \tag{4}$$

where $\lambda \sim \text{Beta}(\alpha, \alpha)$. Third, we used ImageNet-pretrained models for all baseline models to leverage transfer learning, allowing the model to benefit from knowledge learned from a large and diverse image dataset, especially in our setting with limited training data available in medical institutions. Lastly, we applied comprehensive regularization techniques, including label smoothing (Szegedy et al., 2016), learning rate warmup, cooldown (Kalra & Barkeshli, 2024), decay (Smith, 2017), and gradient clipping (Zhang et al., 2020), to enhance the model's reliability and performance.

3.4.2. Confidence calibration

Despite carefully selecting the model architecture and training strategy to address the overconfidence issue, classification models still tend to over-confidently predict AD and HP, as shown in Fig. 4. To mitigate this, we implemented temperature scaling (Hinton et al., 2015), a method that calibrates the model output probabilities to better reflect actual uncertainties, reducing the overconfidence in the predictions. The

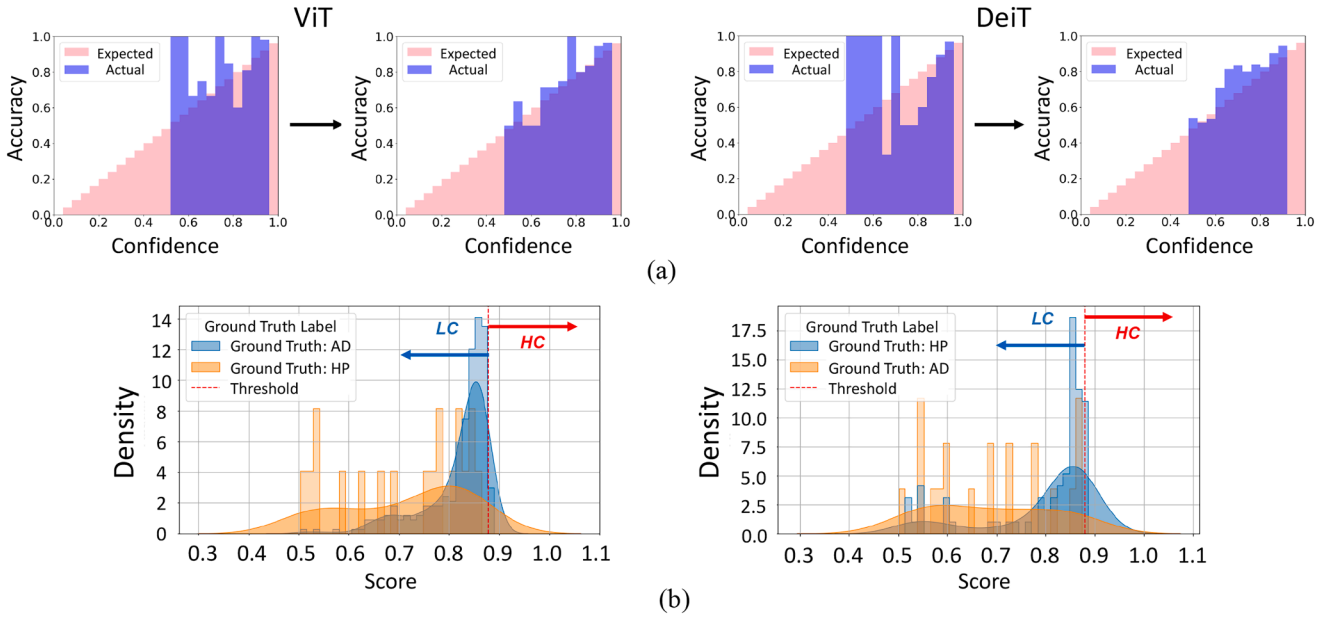


Fig. 4. Training strategies of polyp classification. (a) Reliability plot on the effect of temperature scaling for transformer-based models (Left) Before the temperature scaling, (Right) After the temperature scaling, (b) Setting of the threshold on the uncertainty score distribution into high confidence (HC) and low confidence (LC) of data points as adenomatous polyp (left) and hyperplastic polyp (right).

temperature scaling adjusts the logit vector $\mathbf{z}_i$ by dividing it by a scalar hyperparameter, T, as shown below.

$$\hat{q}_i = \max \sigma\left(\frac{\mathbf{z}_i}{T}\right) \tag{5}$$

Here, $\hat{q}_i$ is the calibrated probability. After applying temperature scaling, the model became more calibrated, effectively addressing the issue of over-confidence.

3.4.3. Polyp uncertainty quantification

Establishing appropriate thresholds for high confidence (HC) and low confidence (LC) classifications is crucial for enhancing the model's performance and interpretability. This ensures that the model's predictions are accurate and reliable, enabling clinicians to make well-informed decisions based on the model's outputs.

Several methods exist for setting the HC and LC thresholds. The first approach involves identifying an optimal threshold that minimizes the difference between the false positive rate (FPR) and actual positive rate (TPR), balancing sensitivity and specificity to distinguish between AD and HP polyps effectively. The second approach focuses on minimizing overall detection error by setting a threshold that reduces the number of misclassifications and optimizing the detection function to improve classification accuracy. The third approach employs expected calibration error (ECE) to ensure the model's confidence scores are well-calibrated (Guo et al., 2017). This method adjusts the threshold so that the predicted confidence scores accurately reflect the true likelihood of the classifications, mitigating both over-confidence and under-confidence in the model's predictions.

In this study, we adopted the third approach–utilizing confidence scores to distinguish between high-confidence (HC) and low-confidence (LC) predictions without interfering with the classification process itself. In medical applications, particularly polyp classification, minimizing false positives, false negatives, and misinterpretation is critical to achieving optimal patient outcomes. Rather than relying solely on accuracy, which provides a binary assessment, our focus expands to include calibration error–a continuous measure ranging from 0 to 1–that more effectively captures the model's confidence reliability and potential for misjudgment.

To determine appropriate thresholds, predicted probabilities were aligned with actual labels and grouped into bins. Candidate thresholds from 0 to 1 were then iteratively evaluated to identify values that minimized the Expected Calibration Error (ECE). As a result, the optimal high-confidence thresholds for AD and HP were identified as 0.899 and 0.888, respectively, rounded to three significant figures.

The classification model and confidence calibration process were designed through comprehensive analysis of model behavior and the adoption of multiple conservative strategies. These efforts aimed to minimize misclassification risk and prevent error accumulation, thereby improving both reliability and clinical usability.

By combining these thresholding techniques with detailed visual analysis, the model achieves high accuracy and confidence calibration. This enhances its clinical applicability by providing endoscopists with a robust tool for precise colorectal polyp classification. Ultimately, the system not only supports accurate diagnoses but also strengthens clinical decision-making for improved patient outcomes.

3.5. Out-of-distribution polyp detection

3.5.1. Feature embedding and divergence

Out-of-distribution (OOD) polyp data refers to test samples that originate from a distribution distinct from the training data (Yang et al., 2021). Identifying such samples is critical for ensuring robust and reliable predictions in clinical settings. To address this, OOD detection is formulated as a feature-space, logit, or probability classification problem, leveraging a corresponding function $\phi(x)$ that maps the input x to a latent space. The decision rule is defined based on a score metric S between the input's value and the in-distribution (ID) value distribution.

$$f_{\text{OOD}}(x) = \begin{cases} \text{OOD}, & \text{if } S(\phi(x), \mathcal{F}) < T_{\text{OOD}}, \\ \text{ID}, & \text{otherwise.} \end{cases} \tag{6}$$

Here, $\phi(x)$ is the feature embedding, logit, or probability of the input x. $\mathcal{F}$ represents the set of tensor values for the ID training data. $S(\phi(x), \mathcal{F})$ is a score metric used to measure how far the input distribution deviates from the ID distribution. T_{OOD} is the threshold for determining whether a sample belongs to OOD, calibrated on validation data.

In colorectal polyp diagnosis, robust OOD detection is crucial for identifying polyps that deviate from the typical characteristics of HP or

AD polyps, preventing misclassification, and ensuring appropriate treatment. Designing an effective OOD method is highly relevant to the application's context and requires in-depth analysis.

3.5.2. Out-of-distribution detection methods

OOD detection methods consist of fine-tuning approaches and post-processing techniques. Fine-tuning methods adjust the decision boundaries of the original baseline model to account for the distribution of additional out-of-distribution data. Post-processing methods, on the other hand, simplify training complexity by evaluating the disparity (Mukhoti et al., 2021) between input samples and training data in the feature space (Sun et al., 2021, 2022; Zhang et al., 2023a) , logit values (Wang et al., 2022), and probabilities (Hendrycks et al., 2022; Hendrycks & Gimpel, 2017; Liang et al., 2018). These methods leverage the differences in the model's response to ID and OOD data (Lee et al., 2018). For example, previously unseen data that significantly deviates from the training data's feature density is assigned a low score, indicating it as OOD. In this study, the following well-known OOD detection methods were investigated.

ODIN applies perturbations to input and temperature scaling for increasing sensitivity to detect subtle differences between ID and OOD samples (Liang et al., 2018). **Maximum softmax probability (MSP)** uses the softmax output of the neural network to identify OOD instances. At the same time, ODIN enhances this approach by applying input perturbations to improve detection accuracy (Hendrycks & Gimpel, 2017). **ReAct** modifies the activation values in the penultimate layer by clipping them to a fixed threshold, reducing the influence of extreme activation caused by OOD samples (Sun et al., 2021). **KLM** compares the Kullback-Leibler (KL) divergence between the predicted distribution and a reference distribution to detect significant deviations from OOD samples (Hendrycks et al., 2022). **GEN** combines outputs from multiple generative models to estimate the likelihood of a sample belonging to the training distribution (Liu et al., 2023). **KNN** uses feature embeddings and distance-based metrics to distinguish between ID and OOD instances (Sun et al., 2022). **ViM** calibrates logits using virtual logits generated for each class from softmax outputs, helping separate ID from OOD (Wang et al., 2022). **SHE** captures class-specific feature patterns from the penultimate layer to determine whether the input sample's feature aligns with any class patterns (Zhang et al., 2023a). **OE** trains the baseline model to distinguish additional outlier data labeled as OOD from ID data (Hendrycks et al., 2019). **MixOE** extends OE by blending ID and OOD data using techniques like Mixup (Zhang et al., 2018), creating a smoother and more robust decision boundary between ID and OOD samples (Zhang et al., 2023b).

3.5.3. Ablation study in baseline model and out-of-distribution detection methods

This section explores how the baseline model and OOD methods work against each polyp types. Fig. 5(b) presents a U-map of the feature, logit, and probability spaces from baseline model for classifying AD, HP, and OOD polyps. The logit values and probabilities of OOD polyps closely align with those of AD and HP polyps, complicating the calibration process to determine whether a sample belongs to the ID or OOD categories. In the feature space, OOD polyps are positioned near the space occupied by HP polyps, making it challenging for feature-based OOD methods to differentiate between HP and OOD polyps effectively.

Likewise, Fig. 5(a) shows that HP and OOD polyps are located closely in the feature space, highlighting the necessity of fine-tuning the baseline model rather than relying on a naïve classification model, even if it demonstrates high performance. To effectively distinguish between HP and OOD polyps, the baseline model must be fine-tuned to adjust decision boundaries in the feature space, and an additional mechanism is needed to detect subtle deviations between the two types of polyps.

Post-processing alone is insufficient to distinguish ID from OOD polyps, as the appearance of these polyps is so similar that even extensive data augmentation and regularization techniques are required for accurate classification of AD and HP polyps. The most critical misclassification occurs when OOD polyps are identified as HP polyps, posing a potential cancer risk. The most significant concern is when the system incorrectly predicts cancerous polyps (either AD or OOD) as HP with high confidence. This could lead to the inappropriate discard of polyps and bypassing necessary pathological examination and follow-up treatment.

To address this challenge, we fine-tuned the baseline model using MixOE, utilizing HP and OOD polyps as training data. This approach allows the model to adjust boundary decisions between HP and OOD polyps without interference from AD polyps. As the number of fine-tuning data is insufficient, the addition of AD polyps interrupted

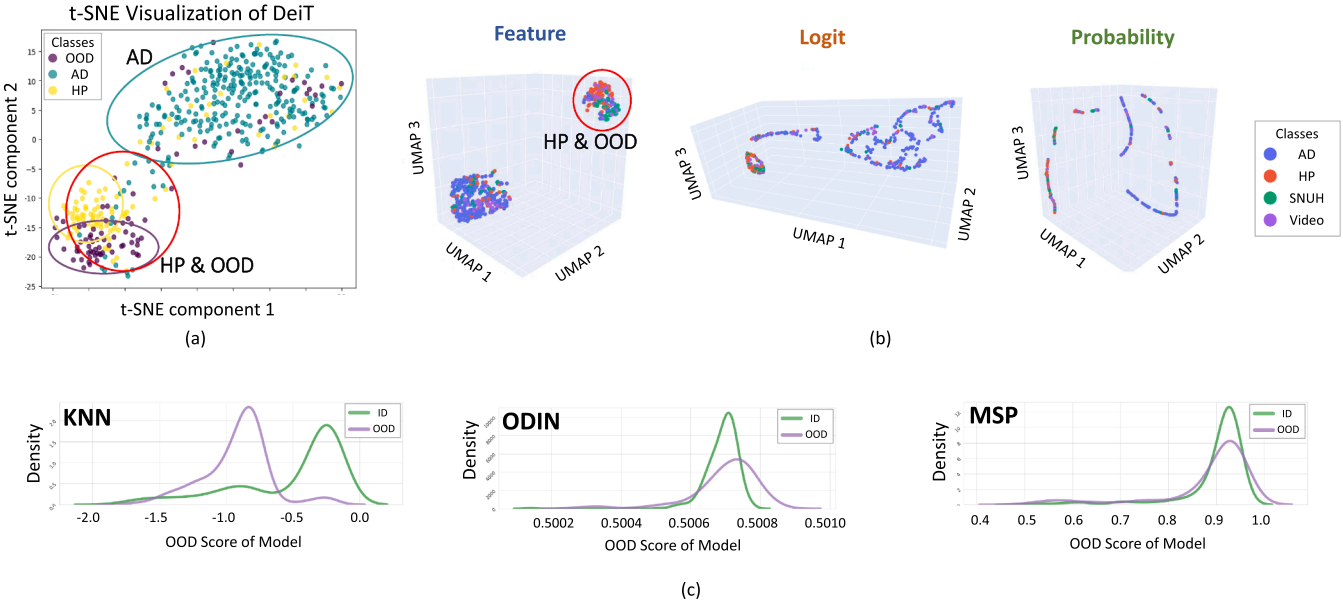


Fig. 5. Preliminary study in three representative out-of-distribution detection methods. (a) t-SNE visualization of tensors of baseline model visualizing the datapoint into the 2D map (b) U-map of feature tensors, logit tensors, probability tensors of baseline model (c) Distribution of OOD scores of representative feature-, logit-, and probability-based methods.

adjusting the decision boundary of the model to distinguish AD, HP, and OOD.

The baseline classification model was fine-tuned with advanced data augmentation techniques like CutMix. We opted for feature-based techniques for the post-processing method, leveraging transfer learning models with complex decision boundaries. The U-map and OOD distribution indicate that logit values and probabilities alone are insufficient to distinguish between ID and OOD polyps. In summary, the OOD detection step employs a fine-tuned MixOE model with HP and OOD polyps combined with a feature-based OOD method.

4. Experimental setup

4.1. Preprocessing

The candidate backbone models are ImageNet-pretrained ViT-S, DeiT-S, and MPViT-S. Images were resized to 256×256 pixels and center-cropped to 224×224 pixels to maintain resolution after polyp localization. The training process incorporated several strategies to address the over-confidence issue posed by the limited amount of data. We augmented the training set using Mixup (Zhang et al., 2018) with $\beta = 0.8$, along with simple augmentations like random horizontal flips with a 0.5 probability and random rotation within a range of $-10°$ to $10°$, to artificially increase data diversity and prevent overfitting. The baseline model has been fine-tuned using MixOE with 10 % of OOD data from Table 2.

4.2. Evaluation metrics

The system's performance is evaluated using multiple metrics, including overall accuracy, which reflects both polyp classification and OOD detection accuracy. Classification and OOD detection accuracy are also reported separately to provide deeper insights into the system's performance.

$$A = \frac{N_{OOD} + N_{ID \cap AD} + N_{ID \cap HP}}{N} \tag{7}$$

where

A : Overall Accuracy,
N_{OOD} : Number of out-of-distribution samples correctly classified,
$N_{ID \cap AD}$: Number of in-distribution samples correctly classified as AD,
$N_{ID \cap HP}$: Number of in-distribution samples correctly classified as HP,
N : Total number of samples.

The expected calibration error (ECE) is measured to assess the calibration error in the models' confidence scores. Additionally, metrics such as the Area Under the Receiver Operating Characteristic curve (AUROC), the Area Under the Precision-Recall curve (AUPRC), the False Positive Rate (FPR) at 0.95 True Positive Rate (TPR) are used to evaluate the quality of OOD detection. These metrics provide a comprehensive assessment of the system's ability to accurately classify polyps and detect OOD instances, ensuring reliable and interpretable results for clinical practice.

4.3. Implementation details

The baseline models were trained using four NVIDIA RTX 3090 GPUs with 20 GB of RAM each. A NIVID A100 GPU with 80 GB of RAM was used for experiments involving several OOD methods requiring higher memory capacity. The models were implemented in Python (v 3.8) and PyTorch (v 1.10.2), and various OOD methods were explored using OpenOOD v1.5 (Zhang et al., 2023c). All evaluations of the full pipeline were performed on a single NVIDIA RTX 3090 GPU with 20 GB of RAM. The implementation is available at https://github.com/sehyunpark99/ColonOOD. The demo can be seen in (Fig.A.1).

Table 3
Comprehensive analysis OOD detection models' performance.

Fine-tuning dataset(Prior classification)	AD, HP and OOD (X)	AD, HP and OOD (O)	HP and OOD (X)	HP and OOD (O)
Center A	27.29 %	49.56 %	69.87 %	**72.49**%
Center A (Video)	27.25 %	49.01 %	64.40 %	**71.95**%
SUN	12.93 %	41.16 %	71.77 %	**79.68**%
POLAR	46.19 %	62.72 %	**75.53**%	74.55 %
CIFAR10	96.82 %	97.85 %	**98.99**%	**98.98**%

5. Results

5.1. Result of ColonOOD framework

By fine-tuning the OOD model on a dataset specifically curated to include a variety of OOD polyps, we achieved improved detection accuracy and better generalization to unseen data. Table 3 presents a comparative analysis of the OOD detection model's performance when trained on two different polyp-type subsets: (AD, HP, and OOD) versus (HP and OOD). The uncertainty-aware classification model classifies AD polyps with high confidence as in-distribution (ID), providing a precedent step for OOD detection. The results show significant improvement, with detection accuracy increasing from 27.29 % to 49.56 % on the Center A dataset and from 12.93 % to 41.16 % on the SUN dataset, demonstrating the effectiveness of leveraging earlier classification steps to boost OOD detection accuracy. However, incorporating AD polyps during fine-tuning degraded the model's performance and accuracy in predicting OOD polyps.

Fine-tuning on HP and OOD samples alone led to substantial improvements, particularly on the POLAR dataset, where performance reached 62.72 %. This suggests that focusing on specific polyp types can significantly enhance OOD detection. The iterative approach, combined with fine-tuning on HP and OOD samples, proved to be the most effective, delivering the best results across all datasets. Notable improvements, such as Center A's increase to 72.49 %, highlight the method's ability to enhance OOD detection through iterative refinement and targeted fine-tuning. The best performance across various datasets demonstrates the pipeline's robustness and effectiveness in practical applications. These findings emphasize the importance of iterative refinement and targeted fine-tuning in developing reliable OOD detection systems for clinical use.

The experimental results show significant improvements in both classification accuracy and OOD detection. The integrated system accurately distinguishes between HP and AD polyps while effectively identifying and managing OOD polyps. Including uncertainty quantification further enhances the reliability of the model's predictions, providing endoscopists with valuable information to support their diagnostic decisions. These results underscore the potential of the proposed system to improve the quality and consistency of colorectal polyp diagnosis in clinical practice.

5.2. Result of uncertainty-aware polyp classification

Table 4 demonstrates that our training strategy outperformed all four baseline models. Each model was deployed in four medical institutes,

Table 4
Performance of different baseline models across various datasets.

	ViT-S	DeiT-S	MP-ViT-S	ResNet-50	Overall
Center A	90.6 %	87.9 %	91.7 %	87.9 %	89.5 %
Center B	89.3 %	86.3 %	86.7 %	86.0 %	87.1 %
Center C	82.4 %	81.6 %	82.8 %	77.9 %	81.2 %
Center D	71.5 %	72.9 %	75.0 %	74.3 %	73.4 %

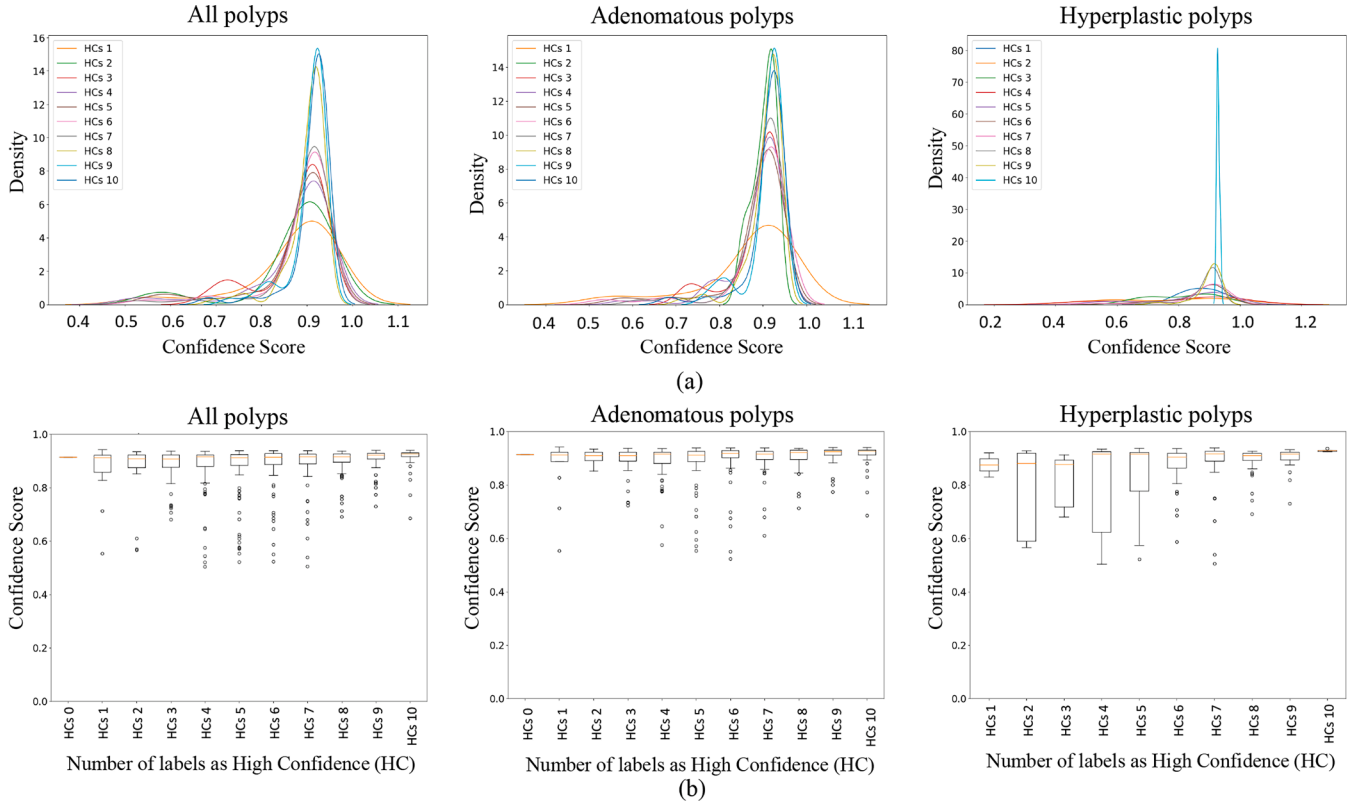


Fig. 6. Distribution of confidence scores across the validation set. Top: Kernel Density Estimate (KDE) of confidence scores, with each line representing the number of endoscopists who classified the polyp as high confidence (HC). Bottom: Box plot showing the number of endoscopists who classified the polyp as HP.

consistently achieving high accuracy. This uniform performance across various settings highlights the robustness and generalizability of our approach.

One notable result is that the outcomes of the uncertainty-aware classification model closely align with expert endoscopists' predictions. The model's confidence scores were compared using an uncertainty assessment dataset, which includes labels indicating high confidence (HC) and low confidence (LC) in polyp optical diagnosis, as determined by ten expert endoscopists. This comparison highlights the model's ability to provide predictions consistent with the judgments of experienced professionals, thereby enhancing its credibility and utility in clinical practice.

Fig. 6 shows that the classification model tends to assign higher confidence scores to polyps that more endoscopists have labeled HC, particularly for HP polyps. This correlation suggests that the model's confidence scores align with the consensus of expert endoscopists, highlighting the model's reliability and potential as a valuable tool in colorectal polyp diagnosis.

5.3. Result of OOD polyp detection

The performance of ViT and DeiT on OOD detection using various methods is detailed in the Appendix. Generally, logit–and probability-based methods poorly differentiate between ID and OOD samples. Methods like MSP, ODIN, and GEN yield over 90 % false positive rates (FPR), indicating their ineffectiveness in distinguishing between ID and OOD samples. The complete performance of each OOD methods is in Figs. B.1 B.2. Our experiments show that feature-based OOD detection methods, particularly KNN combined with fine-tuned models, offer the best performance distinguishing OOD polyps from ID polyps, as hypothesized in the Methods section through our feature, logit, and probability spaces analysis. Feature-based methods leverage the underlying feature representations of polyps, providing a more robust and reliable approach to OOD detection compared to probability-based methods. This confirms the efficacy of feature-based methods for OOD detection and supports their integration into our CAD pipeline.

6. Discussion

The ColonOOD system comprises three main components: automated and manual localization, an uncertainty-aware classification model, and a scenario-targeted OOD detection module that identifies deviations in the feature embeddings of data. The mutually exclusive localization options ensure the system's adaptability across diverse scenarios. The uncertainty-aware classification model is rigorously trained using advanced techniques to enhance robustness and ensure reliable performance in varied clinical environments. Additionally, the novel OOD detection module, consisting of an HP and OOD fine-tuned model and a feature embedding-based post-OOD method, fully leverages the strengths of the robust classification model to mitigate the risk of adverse outcomes through user scenario analysis. Ultimately, ColonOOD provides accurate classification predictions and well-calibrated uncertainty scores, ensuring its results' interpretability, reliability, and clinical utility.

ColonOOD overcomes several limitations of existing CAD systems by synergistically integrating classification and OOD detection models. This approach enables precise classification of AD and HP polyps while effectively identifying and managing OOD polyps, even under covariate shifts in data and clinical environments. The system's uncertainty-aware design offers interpretable predictions, allowing endoscopists to gain actionable insights and make confident, informed decisions. The robust generalizability of ColonOOD, validated across diverse datasets, underscores its potential as a practical solution for real-world clinical workflows.

Extensive evaluations on datasets from private and public medical centers confirm the feasibility and robustness of ColonOOD in diverse clinical settings. ColonOOD aims to improve diagnostic accuracy and reduce reliance on subjective judgment by supporting endoscopists with confidence scores and actionable insights. Testing on colonoscopy videos further demonstrates its reliability by delivering accurate classification results and confidence levels. The complete pipeline achieved an average classification accuracy of 74.67 % for AD, HP, and OOD polyps across four datasets from both private and public sources.

Despite its promising results, ColonOOD has three key limitations that warrant further investigation. First, the limited and non-diverse OOD polyp datasets constrain its evaluation, as the datasets from multiple medical institutions lack sufficient coverage of the numerous types of OOD polyps found worldwide. Expanding OOD data sets to include a wider range of polyp types is crucial to improve the generalizability of the model and ensure robust performance in diverse scenarios.

Second, the system requires more extensive testing in real clinical environments; while ColonOOD has been tested on real colonoscopy videos from a medical center to simulate deployment during procedures, this controlled simulation may not fully capture real-time deployment or interactions with endoscopists. Prospective clinical trials in real-world settings are essential to validate its robustness, usability, and feasibility for practical application.

Finally, the system's polyp localization component may face challenges under environmental variations, such as differences in lighting conditions or image quality during colonoscopy. Although ColonOOD includes both automated and manual localization capabilities, manual localization with human control is currently preferred, as automated localization often fails under certain conditions. Enhancing the automated localization model through adaptive learning techniques and more diverse training data is necessary to ensure consistent performance in various clinical settings.

For further development, ColonOOD can incorporate a feedback loop in the final stage of the system to reduce the risk of edge-case misclassifications. The current system may misclassify borderline OOD samples as AD. Although this scenario does not pose a serious risk, since both categories warrant additional histopathological evaluation, a more advanced system would improve reliability by providing more accurate and targeted classifications.

7. Conclusion

This study presents ColonOOD, a comprehensive and robust CAD system that detects both in-distribution and OOD polyps while providing interpretable uncertainty measures. By effectively addressing data shifts, inconsistent performance, and user-centered design, ColonOOD significantly advances colorectal polyp diagnosis. Its reliable, adaptable, and fully integrated design aligns with clinical workflows, empowering endoscopists with actionable insights to enhance accuracy and efficiency in colorectal cancer prevention and care.

Conflict of interest statement

None of the authors have any conflict of interest.

Data Statement

This study was approved by the Institutional Review Board (IRB) of the following participating centers: Seoul National University Hospital (IRB number: 2105-176-1221), Asan Medical Center (IRB number: 2021-1531), Ewha Medical Center (IRB number: 2021-07-036), and Gangnam Severance Hospital (IRB number: 3-2021-0282). All procedures involving human participants were conducted in accordance with the 1964 Declaration of Helsinki and its later amendments or comparable ethical standards. The SUN dataset was provided by Showa University Northern Yokohama Hospital and Mori Lab, Graduate School of Informatics, Nagoya University. The POLAR dataset was provided by the POLAR study group (https://www.polar.amsterdamumc.org).

CRediT authorship contribution statement

Sehyun Park: Writing – original draft, review & editing, Conceptualization, Investigation, Methodology, Formal Analysis, Validation, Data curation, Visualization, Conceptualization, Experiments, Software. **Dongheon Lee:** Writing – review & editing, Visualization, Supervision. **Ji Young Lee:** Data curation. **Jaeyoung Chun:** Data curation. **Ji Young Chang:** Data curation. **Eunsu Baek:** Writing – review. **Eun Hyo Jun:** Data curation, Writing – review. **Hyung-Sin Kim:** Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Hyung-Sin Kim reports financial support was provided by National Research Foundation of Korea. Sehyun Park reports financial support was provided by National Research Foundation of Korea. Eunsu Baek reports financial support was provided by National Research Foundation of Korea. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This research was supported partly by Creative-Pioneering Researchers Program through Seoul National University, partly by AI-Bio Research Grant through Seoul National University, and partly by the National Research Foundation (NRF) of Korea grant funded by the Korean government (MSIT) (No. RS-2023-00212780, RS-2023-00222663, 2021R1I1A3047535).

Appendix A. Demo

To fully evaluate the pipeline's viability in clinical settings, we tested the entire pipeline using colonoscopy video data and developed a testbed using Streamlit. Operating the full system–including Streamlit, polyp localization, and classification–requires one RTX 3090 GPU and three CPU cores of an Intel Xeon Gold processor, consuming up to 3,000 MB of memory, which is manageable by computers used in colonoscopy procedures. When the clinician pauses on a frame with a polyp and selects the top-left and bottom-left corners, the system crops the polyp into a bounding box for classification. The clinician can then click "classify" to obtain a result, categorized as AD, HP, or OOD, along with the prediction confidence. The entire process takes only 1–2 s. During testing, several issues were identified. First, the automatic localization did not function properly in some cases, indicating that it is not robust enough to handle environmental shifts, such as in external clinical centers. Therefore, we recommended using manual localization to ensure consistent performance alongside the classification model.

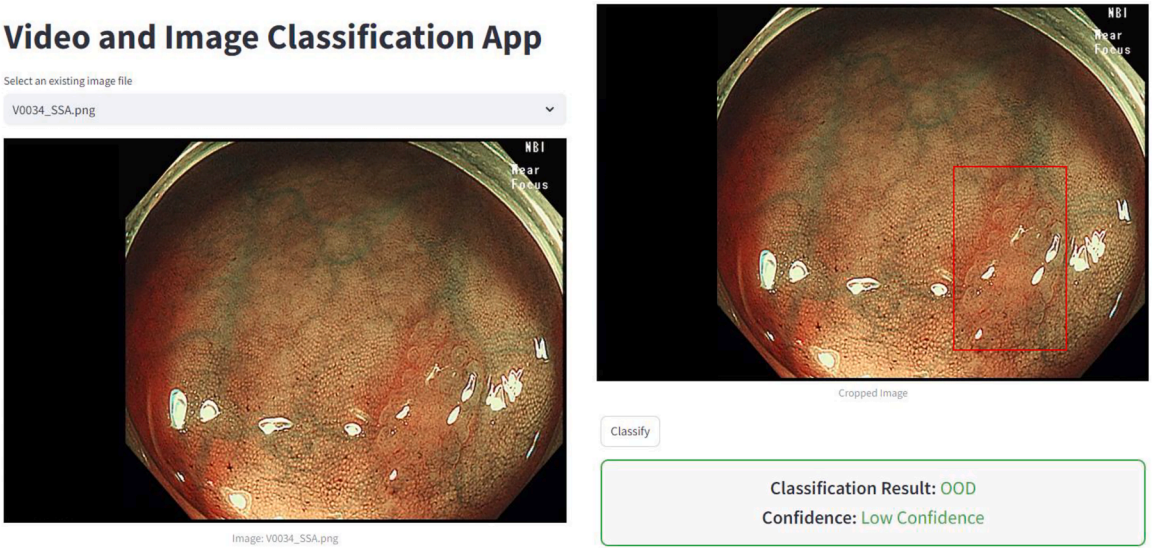


Fig. A.1. Demo.

Appendix B. Appendix Table

(Figs. B.1 B.2).

DeiT						ViT					
Method	Dataset	FPR@95 (↓)	AUROC (↑)	AUPR _IN (↑)	AUPR_OUT (↑)	Method	Dataset	FPR@95 (↓)	AUROC (↑)	AUPR _IN (↑)	AUPR_OUT (↑)
MSP	SNUH	93.82	51.83	81.12	20.5	MSP	SNUH	97.58	53.53	81.65	21.66
	VIDEO	98.92	41.82	76.93	16.74		VIDEO	98.92	45.2	77.15	17.84
	SUN	99.19	42.28	97.23	1.73		SUN	91.13	43.82	98.05	1.56
	POLAR	99.46	48.22	55.29	43.52		POLAR	98.92	38.07	51.86	33.79
	CIFAR10	95.43	58.65	4.19	97.46		CIFAR10	66.13	79.83	13.51	98.7
ODIN	SNUH	98.66	47.14	77.69	18.96	ODIN	SNUH	97.85	44.1	77.77	17.35
	VIDEO	98.66	38.75	75	14.97		VIDEO	98.12	32.66	74.32	12.67
	SUN	96.77	36.33	97.25	1.33		SUN	97.04	28.8	96.9	1.17
	POLAR	100	38.81	49.81	40.94		POLAR	99.73	24.29	44.81	29.57
	CIFAR10	100	41.29	2.74	96.23		CIFAR10	99.73	49.57	3.18	97.17
GEN	SNUH	93.82	51.83	81.12	20.5	GEN	SNUH	97.58	53.53	81.65	21.66
	VIDEO	98.92	41.82	76.93	16.74		VIDEO	98.92	45.2	77.15	17.84
	SUN	99.19	42.28	97.23	1.73		SUN	91.13	43.82	98.05	1.56
	POLAR	99.46	48.22	55.29	43.52		POLAR	98.92	38.07	51.86	33.79
	CIFAR10	95.43	58.65	4.19	97.46		CIFAR10	66.13	79.83	13.51	98.7
ViM	SNUH	69.62	67.37	90.13	27.71	ViM	SNUH	98.92	48.97	78.26	18.64
	VIDEO	73.12	67.82	89.68	26.74		VIDEO	79.84	65.33	89.13	28.27
	SUN	30.38	83.1	99.64	5.34		SUN	41.94	82.87	99.61	21.94
	POLAR	54.3	72.53	84.92	50.84		POLAR	87.9	50.09	65.67	37.62
	CIFAR10	17.2	93.38	81.68	99.65		CIFAR10	1.08	99.72	95.55	99.99
React	SNUH	97.58	54.5	81.55	21.12	React	SNUH	97.31	49.5	78.07	23.37
	VIDEO	98.92	50.71	79.2	20.26		VIDEO	96.24	51.53	80.56	22.07
	SUN	77.69	58.95	98.84	2.59		SUN	100	57.99	97.55	3.01
	POLAR	99.19	46.8	54.82	39.26		POLAR	100	33.75	48.25	32.98
	CIFAR10	48.92	80.18	28.06	98.7		CIFAR10	70.43	83.41	13.5	99.18
KNN	SNUH	69.09	78.44	93.16	36.72	KNN	SNUH	75.81	65.24	89.58	25.72
	VIDEO	76.08	66.41	89.63	26.28		VIDEO	84.95	64.95	89.12	25.91
	SUN	24.19	86.14	99.71	8.24		SUN	37.9	79.38	99.54	20.42
	POLAR	27.42	85.88	92.39	67.53		POLAR	54.84	66.71	81.71	47.04
	CIFAR10	98.95	12.45	97.47	0.48		CIFAR10	9.95	97.95	88.39	99.91
SHE	SNUH	85.48	68.61	90.43	25.92	SHE	SNUH	74.19	75.43	92.97	32.75
	VIDEO	99.19	53.76	79.73	21.49		VIDEO	98.12	56.37	81.26	24.81
	SUN	64.25	64.9	99.11	2.75		SUN	67.74	72.77	99.32	3.46
	POLAR	25.81	86.34	91.21	69.86		POLAR	35.75	79.26	89.03	59.07
	CIFAR10	12.37	94.52	80.46	99.66		CIFAR10	33.06	92.19	54.1	99.61
MDS	SNUH	60.75	69.23	91.35	28.09	MDS	SNUH	98.12	48.28	78.35	18.14
	VIDEO	66.4	69.78	90.88	28.05		VIDEO	73.66	66.38	90.03	27.37
	SUN	28.23	83.18	99.65	5.2		SUN	39.52	84.22	99.65	20.51
	POLAR	46.77	73.23	85.67	51.11		POLAR	83.6	51.55	67.84	38.1
	CIFAR10	17.2	93.01	81.37	99.61		CIFAR10	1.88	99.66	95.05	99.99
KLM	SNUH	97.04	60.55	85.29	23.86	KLM	SNUH	96.77	49.5	79.55	20.3
	VIDEO	90.05	61.09	86.95	22.66		VIDEO	94.62	53.5	83.13	20.35
	SUN	88.44	65.71	98.94	2.8		SUN	82.53	48.08	98.4	1.62
	POLAR	82.53	70.06	77.63	55.51		POLAR	93.01	43.39	58.66	35.58
	CIFAR10	91.94	65.85	5.79	97.85		CIFAR10	95.7	67.04	5.08	98.13

Fig. B.1. OOD Detection Performance of Various OOD Methods.

DeiT				ViT			
Method	**Dataset**	**Accuracy (↑)**	**OOD Accuracy (↑)**	**Method**	**Dataset**	**Accuracy (↑)**	**OOD Accuracy (↑)**
MSP	SNUH	37.99%	39.30%	MSP	SNUH	33.41%	34.06%
	Video	34.07%	35.38%		Video	31.21%	31.87%
	SUN	31.40%	32.98%		SUN	24.01%	24.80%
	POLAR	42.30%	43.27%		POLAR	37.12%	37.60%
	CIFAR10	71.99%	72.05%		CIFAR10	93.64%	93.67%
ODIN	SNUH	18.78%	18.78%	ODIN	SNUH	18.78%	18.78%
	Video	18.24%	18.24%		Video	18.24%	18.24%
	SUN	1.85%	1.85%		SUN	1.85%	1.85%
	POLAR	39.71%	39.71%		POLAR	39.71%	39.71%
	CIFAR10	96.41%	96.41%		CIFAR10	96.41%	96.41%
GEN	SNUH	70.31%	81.22%	GEN	SNUH	72.71%	81.22%
	Video	70.77%	81.76%		Video	73.19%	81.76%
	SUN	84.96%	98.15%		SUN	87.86%	98.15%
	POLAR	52.19%	60.29%		POLAR	53.97%	60.29%
	CIFAR10	3.10%	3.59%		CIFAR10	3.21%	3.59%
ViM	SNUH	48.91%	50.00%	ViM	SNUH	30.79%	31.00%
	Video	47.47%	48.57%		Video	33.63%	33.85%
	SUN	40.90%	42.22%		SUN	21.64%	21.90%
	POLAR	62.88%	63.70%		POLAR	46.84%	47.00%
	CIFAR10	97.84%	97.89%		CIFAR10	97.14%	97.15%
REACT	SNUH	45.63%	47.60%	REACT	SNUH	72.71%	77.51%
	Video	42.86%	44.84%		Video	70.55%	75.38%
	SUN	40.63%	43.01%		SUN	82.32%	88.13%
	POLAR	44.89%	46.35%		POLAR	54.29%	57.86%
	CIFAR10	95.38%	95.47%		CIFAR10	55.63%	55.84%
KNN	SNUH	61.79%	64.41%	KNN	SNUH	37.34%	63.54%
	Video	57.58%	60.22%		Video	36.92%	63.30%
	SUN	56.73%	59.89%		SUN	36.15%	67.81%
	POLAR	72.12%	74.07%		POLAR	47.49%	66.94%
	CIFAR10	98.42%	98.53%		CIFAR10	70.38%	71.54%
SHE	SNUH	69.87%	75.76%	SHE	SNUH	66.16%	68.78%
	Video	64.62%	70.55%		Video	60.00%	62.64%
	SUN	70.45%	77.57%		SUN	62.27%	65.44%
	POLAR	77.63%	82.01%		POLAR	73.91%	75.85%
	CIFAR10	96.86%	97.12%		CIFAR10	94.51%	94.63%
MDS	SNUH	61.35%	64.19%	MDS	SNUH	38.21%	38.65%
	Video	61.54%	64.40%		Video	42.42%	42.86%
	SUN	60.42%	63.85%		SUN	32.72%	33.25%
	POLAR	67.42%	69.53%		POLAR	50.24%	50.57%
	CIFAR10	98.55%	98.68%		CIFAR10	97.54%	97.56%
KLM	SNUH	65.94%	72.05%	KLM	SNUH	72.71%	77.95%
	Video	65.27%	71.43%		Video	71.65%	76.92%
	SUN	73.88%	81.27%		SUN	82.85%	89.18%
	POLAR	60.62%	65.15%		POLAR	54.13%	58.02%
	CIFAR10	37.02%	37.29%		CIFAR10	41.82%	42.06%

Fig. B.2. Overall Accuracy and OOD Accuracy of OOD Methods of DeiT and ViT various datasets.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.eswa.2025.128756

References

POL (2020). Diagnostic performance of a convolutional neural network for diminutive colorectal polyp recognition (POLAR). https://ctv.veeva.com/study/diagnostic-performance-of-a-convolutional-neural-network-for-diminutive-colorectal-polyp-recognition. Clinical Trial Identifier: NCT03822390.

Beger, H., Meining, A., & Shah, M. (2021). Robustness of artificial intelligence systems in gastroenterology: Limitations and future directions. *Clinical Gastroenterology and Hepatology*, *19*(8), 1604–1611. https://doi.org/10.1016/j.cgh.2020.12.003

Chan, E. Y., & Siegel, P. S. (2018). Convolutional neural networks: An overview and application in radiology. *Insights into Imaging*, *9*(3), 391–402. https://doi.org/10.1007/s13244-018-0639-9

Chen, M., Decary, M., & Li, Y. (2021). Artificial intelligence in healthcare: Challenges and opportunities for real-world deployment. *Frontiers in Medicine*, *8*, 650307. https://doi.org/10.3389/fmed.2021.650307

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. *Proceedings of the 34th international Conference on Machine Learning (ICML)*. https://arxiv.org/abs/1706.04599.

Hendrycks, D., Basart, S., Mazeika, M., Zou, A., Kwon, J., Mostajabi, M., Steinhardt, J., & Song, D. (2022). Scaling out-of-distribution detection for real-world settings. *162*, 8759–8773.

Hendrycks, D., & Gimpel, K. (2017). A baseline for detecting misclassified and out-of-distribution examples in neural networks. *International Conference on Learning Representations*. https://openreview.net/forum?id=Hkg4TI9xl.

Hendrycks, D., Mazeika, M., & Dietterich, T. (2019). Deep anomaly detection with outlier exposure. In *International conference on learning representations*.

Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531.

Hong, Z., Yue, Y., Chen, Y., Cong, L., Lin, H., Luo, Y., Wang, M. H., Wang, W., Xu, J., Yang, X., Chen, H., Li, Z., & Xie, S. (2024). Out-of-distribution detection in medical image analysis: A survey. arXiv preprint arXiv:2404.18279.

Itoh, H., Misawa, M., Mori, Y., Oda, M., Kudo, S.-E., & Mori, K. (2020). Sun colonoscopy video database. http://amed8k.sundatabase.org/.

Jin, E.-H., Lee, D.-H., Bae, J. H., Kang, H. Y., Kwak, M. S., Seo, J. Y., Yang, J. I., Yang, S. Y., Lim, S. H., Yim, J. Y., Lim, J. H., Chung, G., Chung, S.-J., Choi, J. M., Han, Y. M., Kang, S. J., Lee, J., Kim, H. C., & Kim, J. S. (2020). Improved accuracy in optical diagnosis of colorectal polyps using convolutional neural networks with visual explanations. *Gastroenterology*, *158*(8), 2169–2179. https://doi.org/10.1053/j.gastro.2020.02.036

Kalra, D. S., & Barkeshli, M. (2024). Why warmup the learning rate? Underlying mechanisms and improvements. arXiv preprint arXiv:2406.09405.

Kiwan, W., Patel, S., Judd, S., Nas, H., Goyal, S., & Antaki, F. (2022). Involvement of gastroenterology fellows in colonoscopy improves the adenoma detection rate: A retrospective cohort study. *Research Square*. License: This work is licensed under a Creative Commons Attribution 4.0 International License. https://doi.org/10.21203/rs.3.rs-40961/v1

Lee, K., Lee, K., Lee, H., & Shin, J. (2018). A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in neural information processing systems 31: Annual conference on neural information processing systems 2018, NeurIPS 2018, December 3–8, 2018, Montréal, Canada* (pp. 7167–7177).

Liang, S., Li, Y., & Srikant, R. (2018). Enhancing the reliability of out-of-distribution image detection in neural networks. https://openreview.net/forum?id=H1VGkIxRZ.

Liu, X., Lochman, Y., & Christopher, Z. (2023). GEN: Pushing the limits of softmax-based out-of-distribution detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

Lo, C.-M., Yeh, Y.-H., Tang, J.-H., Chang, C.-C., & Yeh, H.-J. (2022). Rapid polyp classification in colonoscopy using textural and convolutional features. *Healthcare (Basel)*, *10*(8), 1494. https://doi.org/10.3390/healthcare10081494

Mukhoti, J., Kirsch, A., van Amersfoort, J. R., Torr, P. H. S., & Gal, Y. (2021). Deterministic neural networks with appropriate inductive biases capture epistemic and aleatoric uncertainty. *ICML*, abs/2102.11582.

Ozawa, T., Ishihara, S., Fujishiro, M., Kumagai, Y., Shichijo, S., & Tada, T. (2020). Automated endoscopic detection and classification of colorectal polyps using convolutional neural networks. *Therapeutic Advances in Gastroenterology*, *13*. https://doi.org/10.1177/1756284820910659

Parsa, N., Hassan, C., Paggi, M., Fabbri, C., Pellisé, M., Bisschops, R., & Dekker, E. (2021). Colorectal polyp characterization with standard endoscopy: Will artificial intelligence succeed where human eyes failed? *Best Practice & Research Clinical Gastroenterology*, *52–53*, 101736. https://doi.org/10.1016/j.bpg.2021.101736

Patel, S. G., Scott, F. I., Das, A., Rex, D. K., McGill, S., Kaltenbach, T., Ahnen, D. J., Rastogi, A., & Wani, S. (2020). Cost effectiveness analysis evaluating real-time characterization of diminutive colorectal polyp histology using narrow band imaging (NBI). *Journal of Gastroenterology*, *1*(1), 1–15.

Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C., & Dosovitskiy, A. (2021). Is it time to replace CNNs with transformers for medical images? https://arxiv.org/abs/2108.09038.

Seo, J. Y., Jin, E. H., Bae, J. H., Kim, H. J., Joo, N., Lee, J. M., Kim, Y. H., Park, E., Lee, J. H., Ryu, J. S., Yang, H. J., Oh, D.-H., & Lee, B. S. (2020). Multidirectional colonoscopy quality improvement increases adenoma detection rate: Results of the seoul national university hospital healthcare system gangnam center colonoscopy quality upgrade project (gangnam-CUP). *Digestive Diseases and Sciences*, *65*, 1806–1815. https://doi.org/10.1007/s10620-019-05944-5

Smith, J., Doe, J., & Others (2024). Missed adenomatous polyps and their impact on post-colonoscopy colorectal cancer rates. *BMC Gastroenterology*, *24*(1), 365. https://doi.org/10.1186/s12876-024-03365-x

Smith, L. N. (2017). Cyclical learning rates for training neural networks. arXiv:1506.01186arXiv preprint arXiv:1506.01186.

Sun, Y., Guo, C., & Li, Y. (2021). React: Out-of-distribution detection with rectified activations. *Advances in Neural Information Processing Systems*, *34*, 144–157.

Sun, Y., Ming, Y., Zhu, X., & Li, Y. (2022). Out-of-distribution detection with deep nearest neighbors. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvári, G. Niu, & S. Sabato (Eds.), *International conference on machine learning, ICML 2022, 17–23 July 2022, Baltimore, Maryland, USA* (pp. 20827–20840). PMLR (*vol. 162*). Proceedings of Machine Learning Research.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2818–2826).

Urban, G., Tripathi, P., Alkayali, T., Mittal, M., Jalali, F., Karnes, W., & Samarasena, J. (2018). Deep learning-based detection of colon polyps in colonoscopy images. *Gastroenterology*, *155*(4), 1069–1078. https://doi.org/10.1053/j.gastro.2018.06.037

Wang, H., Li, Z., Feng, L., & Zhang, W. (2022). Vim: Out-of-distribution with virtual-logit matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

Wang, P., Berzin, T. M., Glissen Brown, J. R., Singh, P., Liu, P., Zhou, C., & Xiong, F. (2020). Real-time automatic detection system for colonoscopy: Analysis of 30,000 colonoscopies. *Gastrointestinal Endoscopy*, *91*(2), 470–478. https://doi.org/10.1016/j.gie.2019.09.041

Yang, J., Zhou, K., Li, Y., & Liu, Z. (2021). Generalized out-of-distribution detection: A survey. arXiv preprint arXiv:2110.11334.

Yang, X., Song, E., Ma, G., Zhu, Y., Yu, D., Ding, B., & Wang, X. (2023). YOLO-OB: An improved anchor-free real-time multiscale colon polyp detector in colonoscopy. arXiv preprint arXiv:2312.08628.

Yang, Y. J., Cho, B.-J., Lee, M.-J., Kim, J. H., Lim, H. S., Bang, C. S., Jeong, H. M., Hong, J. T., & Baik, G. H. (2020). Automated classification of colorectal neoplasms in white-light colonoscopy images via deep learning. *Journal of Clinical Medicine*, *9*(5), 1593. https://doi.org/10.3390/jcm9051593

Yoshida, N., Inoue, K., Tomita, Y. et al. (2021). An analysis about the function of a new artificial intelligence, CAD EYE with the lesion recognition and diagnosis for colorectal polyps in clinical practice. *International Journal of Colorectal Disease*, *36*, 2237–2245. https://doi.org/10.1007/s00384-021-04006-5

Zhang, H., Cisse, M., Dauphin, Y. N., & Lopez-Paz, D. (2018). Mixup: Beyond empirical risk minimization. In *International conference on learning representations*.

Zhang, J., Fu, Q., Chen, X., Du, L., Li, Z., Wang, G., Liu, x., Han, S., & Zhang, D. (2023a). Out-of-distribution detection based on in-distribution data patterns memorization with modern hopfield energy. In *The eleventh international conference on learning representations*.

Zhang, J., Inkawhich, N., Linderman, R., Chen, Y., & Li, H. (2023b). Mixture outlier exposure: Towards out-of-distribution detection in fine-grained environments. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision (WACV)* (pp. 5531–5540).

Zhang, J., Yang, J., Wang, P., Wang, H., Lin, Y., Zhang, H., Sun, Y., Du, X., Zhou, K., Zhang, W., Li, Y., Liu, Z., Chen, Y., & Li, H. (2023c). OpenOOD v1.5: Enhanced benchmark for out-of-distribution detection. arXiv preprint arXiv:2501.09298.

Zhang, Z., Bartlett, P. L., Rebollo-Monedero, D., Wu, Y., & Jordan, M. I. (2020). Why gradient clipping accelerates training: A theoretical justification for adaptivity. arXiv preprint arXiv:1905.11881.