Article

Not peer-reviewed version

Real-World Colonoscopy Video Integration to Improve Artificial Intelligence Polyp Detection Performance and Reduce Manual Annotation Labor

Yuna Kim , Ji-Soo Keum , Jie-Hyun Kim * , Jae-Young Chun * , Sang-Il Oh , Kyung-Nam Kim , Young-Hoon Yoon , Hyojin Park

Posted Date: 13 March 2025

doi: 10.20944/preprints202503.0938.v1

Keywords: artificial intelligence; colon polyp; colon cancer; colonoscopy

 doi:10.20944/preprints202503.0938.v1

Article

Real-World Colonoscopy Video Integration to Improve Artificial Intelligence Polyp Detection Performance and Reduce Manual Annotation Labor

Yuna Kim [1,†], Ji-Soo Keum [2,†], Jie-Hyun Kim [1,*], Jae-Young Chun [1,*], Sang-Il Oh [2], Kyung-Nam Kim [2], Young-Hoon Yoon [1] and Hyojin Park [1]

[1] Department of Internal Medicine, Division of Gastroenterology, Gangnam Severance Hospital, Yonsei University College of Medicine, Seoul 06273, Republic of Korea

[2] Waycen Inc., Seoul 06167, Republic of Korea

* Correspondence: Jie-Hyun Kim, MD, PhD; Department of Internal Medicine, Division of Gastroenterology, Gangnam Severance Hospital, Yonsei University College of Medicine 20, Eonju-ro 63-gil, Gangnam-gu, Seoul 06229, Korea; Tel: +82–2-2019-3505; Fax: +82-2-3463-3882; E-mail: otilia94@yuhs.ac; Jaeyoung Chun, MD; Department of Internal Medicine, Division of Gastroenterology, Gangnam Severance Hospital, Yonsei University College of Medicine 20, Eonju-ro 63-gil, Gangnam-gu, Seoul 06229, Korea; Tel: +82–2-2019-4371; Fax: +82-2-3463-3882 Email: chunjmd@yuhs.ac

† Yuna Kim and Ji-Soo Keum contributed equally as the first authors of this study.

Abstract: Background/Objectives: Artificial intelligence (AI) integration in colon polyp detection often exhibits high sensitivity but notably low specificity in real-world settings, primarily due to reliance on publicly available datasets alone. To address this limitation, we proposed a semi-automatic annotation method using real colonoscopy videos to enhance AI model performance and reduce manual labeling labor. **Methods**: An integrated AI model was trained and validated on 86,258 training images and 17,616 validation images. Model 1 utilized only publicly available datasets, while Model 2 additionally incorporated images obtained from real colonoscopy videos of patients through a semi-automatic annotation process, significantly reducing the labeling burden on expert endoscopists. **Results**: The integrated AI model (Model 2) significantly outperformed the public-dataset-only model (Model 1). At epoch 35, Model 2 achieved a sensitivity of 90.6%, specificity of 96.0%, overall accuracy of 94.5%, and an F1 score of 89.9%. All polyps in test videos were successfully detected, demonstrating considerable enhancement in detection performance compared to the public dataset-only model. **Conclusions**: Integrating real-world colonoscopy video data using semi-automatic annotation markedly improved diagnostic accuracy while potentially reducing the need for extensive manual annotation typically performed by expert endoscopists. However, the findings need validation through multicenter external datasets to ensure generalizability.

Keywords: artificial intelligence; colon polyp; colon cancer; colonoscopy

1. Introduction

Artificial intelligence (AI) integration into medical imaging, particularly colonoscopy image analysis, has become increasingly relevant in clinical practice due to its potential to enhance polyp detection and diagnosis [1–3]. AI models supporting colonoscopy procedures generally fall into two categories: Computer-aided detection (CADe), which focuses on detecting abnormalities, such as polyps, and Computer-aided diagnosis (CADx), which classifies detected lesions based on their characteristics [4].

Despite advancements in AI models for colonoscopy, the lack of easily accessible, high-quality data for training and validating AI models remains a challenge [5]. Models trained exclusively on publicly available datasets frequently encounter difficulties due to discrepancies in image quality,

 doi:10.20944/preprints202503.0938.v1

diversity, and differences in clinical equipment, resulting in high sensitivity but notably low specificity in clinical settings [6,7].

This study aimed to address these limitations through a novel semi-automatic annotation method utilizing real patient colonoscopy videos from Gangnam Severance Hospital. Unlike previous methods relying heavily on manual annotation by expert endoscopists, our approach leveraged preliminary AI inference results for efficient selection of clinically relevant video frames, substantially reducing annotation effort. Although semi-automatic annotation is a common method in general computer vision tasks, our specific approach—AI-driven preliminary selection from colonoscopy video data—is designed explicitly for clinical practice integration.

This study demonstrates the feasibility of developing an accurate and efficient AI polyp detection model by integrating real-world clinical data, significantly reducing manual annotation burdens. Nevertheless, due to the single-center nature of the data collection, additional multicenter validation is necessary to ensure broader applicability.

2. Materials and Methods

2.1. Patients

Colonoscopy video data were retrospectively collected from patients who underwent routine colonoscopies at Gangnam Severance Hospital, a tertiary academic center, between April 2021 and April 2022. Inclusion criteria consisted of adults (≥18 years old) undergoing colonoscopy for indications including colorectal cancer screening, surveillance, or evaluation of gastrointestinal symptoms. Videos with poor bowel preparation were excluded. Personally identifiable information was anonymized prior to analysis.

The dataset included 117 colonoscopy videos totaling 30 hours, 57 minutes, and 46 seconds, averaging approximately 15 minutes and 53 seconds per patient. Overall, the dataset comprised 3,348,994 frames, averaging 28,624 frames per video.

2.2. Public Datasets

Representative public datasets used for colonoscopy research include ETIS-Larib [8], CVC-ClinicDB [9], KVASIR-SEG [10], LDPolypVideo [11], KUMC [12], and PolypGen [13]. Each dataset offered unique features and contributions to polyp detection and segmentation tasks: (a) ETIS-Larib Polyp DB contains 196 image frames with binary masks of polyps. (b) CVC-ClinicDB provides 612 images with binary masks extracted from 29 colonoscopy videos. (c) Kvasir-SEG offers 1,000 polyp images with corresponding masks. (d) LDPolypVideo includes 160 labeled and 42 unlabeled videos, containing 33,884 images annotated with bounding boxes. (e) KUMC provides 37,899 frames with bounding box annotations. (f) PolypGen contains 1,537 polyp images, each accompanied by binary masks.

2.3. Data Processing

Colonoscopy videos were used to develop two deep learning-based AI models. The collected colonoscopy videos (n=117) from Gangnam Severance Hospital were randomly allocated into three groups: training (57 videos), validation (15 videos), and test sets (45 videos), using a computer-generated randomization procedure to minimize selection bias. The extracted frames from these videos, combined with publicly available datasets (ETIS-Larib, CVC-ClinicDB, KVASIR-SEG, LDPolypVideo, KUMC, PolypGen), were utilized to construct comprehensive image datasets for model training, validation, and testing. The detailed distribution of images across the datasets is summarized in Table 1.

 doi:10.20944/preprints202503.0938.v1

Table 1. Summary of the number of images used for training, validation and test.

Dataset	Training (polyps)	Validation (polyps)	Test (polyps)	Total (polyps)
A. ETIS-Larib	196 (196)	-	-	196 (196)
B. CVC-ClinicDB	612 (612)	-	-	612 (612)
C. KVASIR-SEG	900 (900)	-	-	900 (900)
D. LDPolypVideo	23,723 (1,604)	6,191 (462)	-	29,914 (2,066)
E. KUMC	27,048 (27,048)	4,214 (4,214)	-	31,262 (31,262)
F. PolypGen	980 (980)	200 (200)	-	1,180 (1,180)
G. Gangnam Severance Hospital	32,799 (9,353)	7,011 (623)	4,373 (1,200)	44,183 (11,176)
Total	86,258 (40,693)	17,616 (5,499)	4,373 (1,200)	108,247 (47,392)

Two deep-learning AI models were developed to compare the effects of integrating real-world clinical data. Model 1 was trained exclusively using publicly available datasets, while Model 2 incorporated additional images obtained from colonoscopy videos collected at Gangnam Severance Hospital. Both models were trained using images containing polyps and normal colonoscopic images without polyps, as outlined in Table 2. The final evaluation was performed using the best-performing model, as determined by accuracy on the validation set.

Table 2. Dataset configuration used to develop AI models.

Model	Dataset	Number of images (polyps)			
		Training	Validation	Test	Total
1	A. + B + C + D + E + F (Public dataset only)	53,459 (31,340)	10,605 (4,876)	4,373 (1,200)	68,437 (37,416)
2	A. + B + C + D + E + F + G (Public + Collected dataset)	86,258 (40,693)	17,616 (5,499)	4,373 (1,200)	108,247 (47,392)

2.4. Semi-Automatic Image Selection

The proposed development process is illustrated in Figure 1. Initially, Model 1 was trained and validated exclusively using public datasets. Subsequently, Model 1 was used to infer frames from 72 colonoscopy videos (57 training, 15 validation) collected from Gangnam Severance Hospital. In this inference step, clinical timestamps indicating the actual removal of polyps during colonoscopy were indirectly used to guide the AI model in identifying frames likely containing polyps. The inference results were automatically categorized into subfolders as frames containing polyps and frames without polyps. Each detected polyp area was highlighted with bounding boxes on corresponding video frames and saved separately, facilitating efficient review by researchers. Frames incorrectly classified by Model 1—such as missed polyps (false negatives) or incorrectly identified polyps (false positives)—were also selectively included to enhance subsequent model training. This combined inference and data-selection approach, termed the semi-automatic image selection process, substantially reduced manual annotation effort, enabling rapid construction of the training dataset for Model 2.

 doi:10.20944/preprints202503.0938.v1

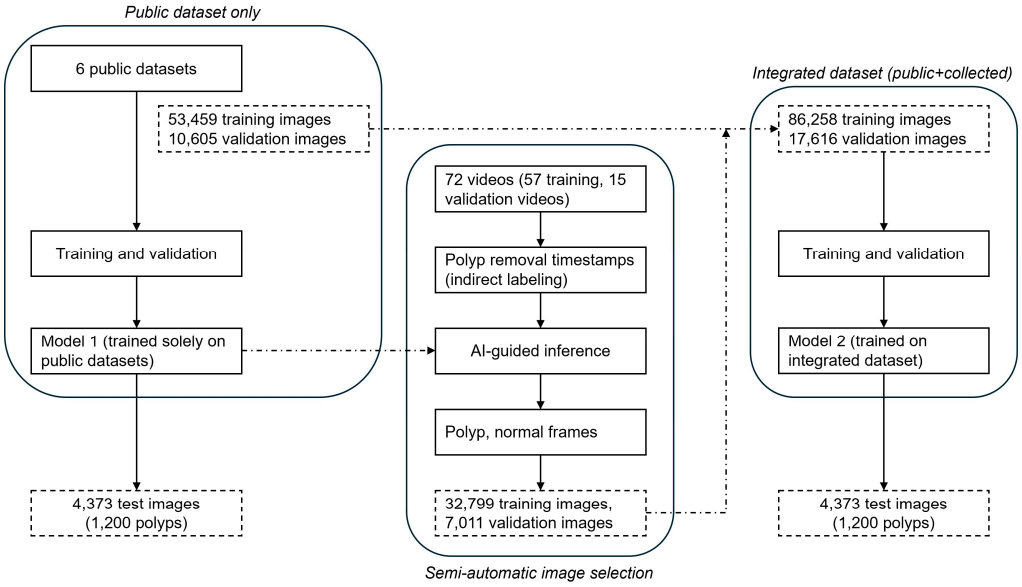

Figure 1. Process of developing the initial model with public date sets and constructing additional datasets.

Figure 2 shows representative examples of inference results from Gangnam Severance Hospital using Model 1 trained on public datasets. Each row represents a sequence of consecutive frames. Frames with blue borders indicate instances where polyps were inaccurately detected or completely missed. These misdetected frames were selectively incorporated into the training dataset to enhance the subsequent learning of Model 2.

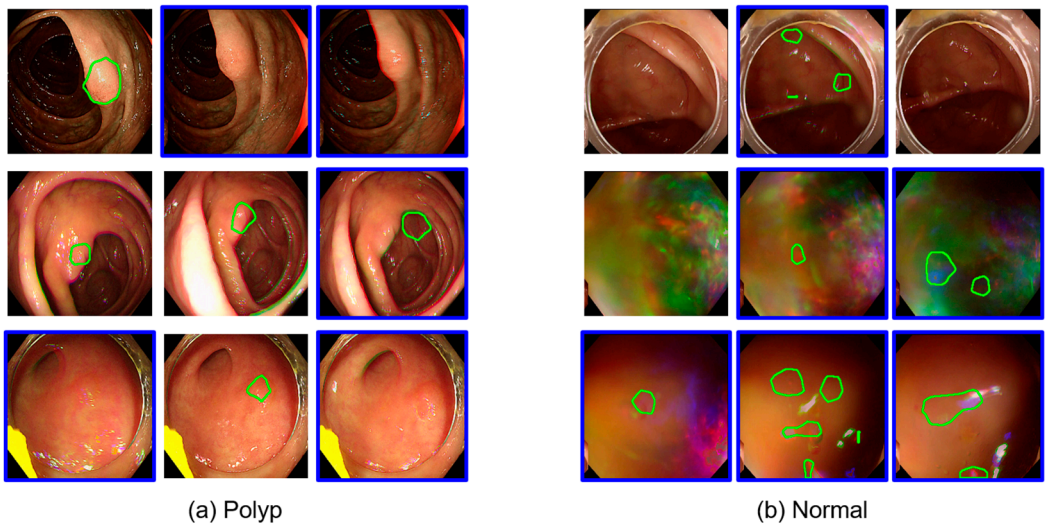

Figure 2. Examples of resulting images from inference in Model 1. The images include detected polyps (a) and normal tissue (b). Blue squares indicate data selected for further training.

2.5. Algorithm Training, Validation, and Test Sets

A YOLO-based object detection model, a deep learning approach, was employed to detect polyps in colonoscopy images. Previous studies have shown that various backbones based on convolutional neural networks (CNNs) exhibit similar performance when applied to endoscopic images [2,3,6,12]. Many existing AI models used for colonoscopy have initially been developed for general image recognition tasks and subsequently adapted for specific object detection tasks, including polyp detection.

 doi:10.20944/preprints202503.0938.v1

In this study, we adopted a CNN-based polyp detection model utilizing the Visual Geometry Group (VGG) backbone. The VGG backbone has been widely applied in medical image analysis and has demonstrated reliable performance in extracting relevant image features in various endoscopic imaging tasks [14,15].

To train Model 1, we augmented the training data (public datasets only) by a factor of 10 using various data augmentation methods, including vertical and horizontal flips, left-right symmetry, random cropping, and rotation. Input image size was standardized at 224 × 224 pixels, and polyp masks and bounding box annotations were consistently preserved during augmentation. Model 1 was trained for 50 epochs with a VGG backbone followed by fully connected layers, optimized using the Adam optimizer with an initial learning rate of 1e-05. Cross-entropy and activation errors served as error functions for backpropagation.

Model 2 integrated both public datasets and additional images derived from colonoscopy videos obtained at Gangnam Severance Hospital. Frames from these videos were efficiently annotated using a semi-automatic selection process based on preliminary inference by Model 1. Model 2 underwent similar training procedures as Model 1, with validation performance assessed at each epoch. The optimal performing model on the validation set was selected for subsequent testing.

The test dataset comprised 4,373 previously unseen images, including 1,200 polyp-containing and 3,173 non-polyp-containing images, extracted from 45 separate test videos. Specifically, 120 polyps were identified in these test videos, and approximately 10 sequential frames per polyp (at 5-frame intervals, ~1.6 seconds duration per polyp) were extracted for performance evaluation.

2.6. Evaluations

Several metrics were used to evaluate the model performance [16,17]. True positives (TP) represent the number of polyps that were accurately detected by the model. The true Negative (TN) indicates the number of normal images correctly classified as normal. A false Negative (FN) refers to instances in which a polyp was present but not detected by the model, whereas a False Positive (FP) denotes cases in which a polyp was incorrectly identified in a normal image.

Sensitivity (or recall) reflects the proportion of actual polyps that were correctly identified by the model and was computed as TP divided by the sum of TP and FN. Specificity represents the proportion of actual normal images that were correctly classified and was calculated as TN divided by the sum of TN and FP. Accuracy measures the overall proportion of correctly classified images, both polyps and normal, and was determined by dividing the sum of TP and TN by the total number of images (TP + TN + FP + FN). Additionally, the F1 Score, which is the harmonic means of sensitivity and positive predictive value (PPV), provides a balanced measure of precision and recall. The F1 Score was calculated as twice the product of sensitivity and PPV divided by the sum of sensitivity and PPV, where PPV was calculated as TP divided by the sum of TP and FP.

3. Results

The performance of the proposed method was evaluated using the model from the epoch that demonstrated the highest sensitivity and specificity of the validation data. This evaluation was performed over 50 epochs for Model 1, which was trained solely with public datasets, and Model 2, which was trained with both public and additional collected data. Model 1 achieved an optimal performance at epoch 37, whereas Model 2 reached its peak performance at epoch 35. Table 3 presents the polyp detection performance on the test data using the model saved at epoch 37 for Model 1 and the models saved at 5-epoch intervals from epoch 5 to epoch 45 for Model 2.

Table 3. Comparison of polyp detection performance between the model trained only with public datasets and the integrated model that includes additional data from colonoscopy videos at Gangnam Severance Hospital.

Model 1	**Model 2**

 doi:10.20944/preprints202503.0938.v1

	Epoch 37	Epoch 5	Epoch 10	Epoch 15	Epoch 20	Epoch 25	Epoch 30	Epoch 35	Epoch 40	Epoch 45
TP	933	948	1,017	1,056	955	895	1,018	1,087	1,051	1,042
FN	267	252	183	144	245	305	182	113	149	158
FP	3,647	537	365	427	308	160	184	132	180	148
TN	230	3,104	3,100	3,100	3,141	3,155	3,149	3,142	3,144	3,144
Sensitivity	0.778	0.790	0.848	0.880	0.796	0.746	0.848	0.906	0.876	0.868
Specificity	0.059	0.853	0.895	0.879	0.911	0.952	0.945	0.960	0.946	0.955
PPV	0.204	0.638	0.736	0.712	0.756	0.848	0.847	0.892	0.854	0.876
F1 score	0.323	0.706	0.788	0.787	0.775	0.794	0.848	0.899	0.865	0.872
Accuracy	0.229	0.837	0.883	0.879	0.881	0.897	0.919	0.945	0.927	0.932

The test results revealed that Model 1, trained only with public datasets, had a sensitivity of 0.778 for detecting polyps, indicating reasonable performance. However, it exhibited a low specificity of 0.059 due to significant false detections, which were attributed to differences in image quality between the collected data and public datasets. In contrast, Model 2, which incorporated additional collected data into the training process, demonstrated an improved performance by addressing these environmental differences. Specifically, Model 2 achieved a sensitivity of 0.906, specificity of 0.960, and accuracy of 0.945 at epoch 35, reflecting a notable enhancement in detection capability.

In real-time colonoscopy applications, the polyp detection model must minimize false detections in normal images and ensure that all polyps are accurately identified to prevent them from being overlooked. Frequent false detections can lead to examiner fatigue and reduce the reliability of the detection model. Conversely, accurate detection of polyps occasionally provides opportunities to identify and remove lesions that might otherwise be missed. Table 4 details the number and detection rates of polyps that were never detected out of the 120 polyps, each represented by 10 images, amounting to 1,200 test images. Figure 3 shows examples of the 120 undetected polyps. Figure 4 shows the number of polyps accurately detected in each epoch. The results indicated that Model 2, which utilized additional collected data, consistently demonstrated higher sensitivity than Model 1, which relied solely on public datasets. This suggests that Model 2 offers an improved continuity of polyp detection throughout the video sequences.

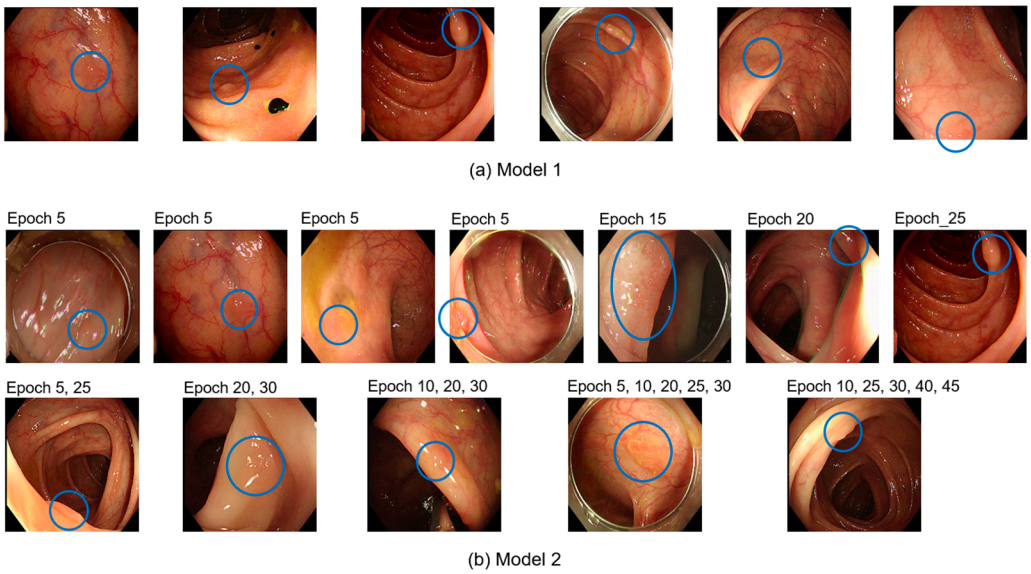

Figure 3. Example of polyps that were not detected by either Model 1 (a) or Model 2 (b) among 120 test polyps.

 doi:10.20944/preprints202503.0938.v1

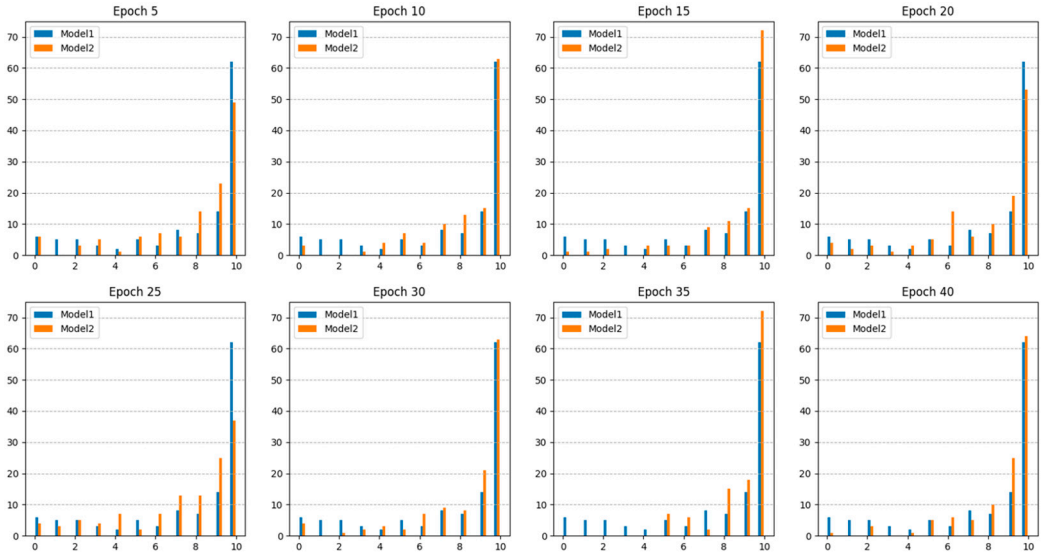

Figure 4. The number of TPs for polyps in each epoch. The X-axis represents the number of TPs, while the Y-axis indicates the number of polyps.

Table 4. Comparison of polyp detection performance between Model 1 and Model 2.

	Model 1	**Model 2**								
	Epoch 37	**Epoch 5**	**Epoch 10**	**Epoch 15**	**Epoch 20**	**Epoch 25**	**Epoch 30**	**Epoch 35**	**Epoch 40**	**Epoch 45**
Undetected polyp	6	6	3	1	4	4	4	0	1	1
Detection rate (%)	95.00	95.00	97.50	99.17	96.67	96.67	96.67	100.0	99.17	99.17

4. Discussion

Recent advances in AI-based CADe and CADx have significantly improved the detection and characterization of colorectal polyps during colonoscopy, potentially enhancing clinical outcomes through earlier and more accurate diagnosis [18,19]. However, AI models trained exclusively on publicly available datasets frequently demonstrate limited specificity when applied to real-world clinical scenarios, mainly due to differences in image quality, diversity, and endoscopic equipment [20,21]

In this study, we proposed a simple and efficient approach to enhance the performance of AI models by integrating real-world colonoscopy video data into the training process. A distinctive feature of our method was the semi-automatic annotation technique, which utilized preliminary inference results from a publicly trained AI model to rapidly and automatically categorize colonoscopy video frames into polyp-containing and non-polyp-containing categories. Although semi-automatic annotation tools, such as CVAT, are widely used in computer vision, our approach specifically employed AI-guided inference to minimize manual intervention by expert endoscopists, thereby significantly reducing the traditionally labor-intensive annotation workload. [22].

By integrating real patient colonoscopy data through this approach, we observed notable improvements in both sensitivity and specificity compared to a model trained solely on public datasets. This highlights that our method not only improves AI model performance but also offers practical advantages by substantially reducing manual annotation efforts.

However, several limitations of our study warrant careful consideration. First, this study used colonoscopy videos from a single institution employing a standardized endoscopic imaging system

 doi:10.20944/preprints202503.0938.v1

(Olympus Evia Exera III CV-190). Therefore, generalizability of our results to other clinical environments and various endoscopic equipment has not yet been confirmed [23]. Future studies should include external validation using multi-center datasets to verify broader clinical applicability and reproducibility of results. Additionally, the study lacked detailed patient characteristics data, including age, gender, clinical indications, polyp morphology, size, location, and histopathological results. Consequently, we could not evaluate the AI model's performance across different lesion types or diverse patient subpopulations, potentially limiting the clinical interpretation of our results. Further research incorporating comprehensive clinical and histopathological data is required to better understand the AI model's performance across various lesion types and patient subpopulations, which will be crucial for identifying potential areas of improvement and enhancing clinical relevance. Furthermore, our evaluation focused primarily on per-frame metrics such as sensitivity and specificity, which may not fully reflect clinical relevance. To better evaluate AI systems in clinical settings, future studies should incorporate lesion-level sensitivity, precision metrics per detected polyp, and precision-recall curves or receiver operating characteristic (ROC) analyses.

In conclusion, our study introduces a practically meaningful, efficient, and scalable approach to significantly enhance AI model performance for colonoscopy polyp detection using real-world colonoscopy video data. The semi-automatic annotation method employed could substantially reduce the traditionally labor-intensive manual annotation process typically performed by expert endoscopists, making it an attractive strategy for practical clinical implementation. However, further external validation, detailed patient and lesion characteristic analyses, and lesion-level performance assessments are necessary to confirm and maximize the clinical utility and generalizability of our approach.

Author Contributions: Conceptualization, Jie-Hyun Kim; Data curation, Yuna Kim, Ji-Soo Keum, Sang-Il Oh and Kyung-Nam Kim; Formal analysis, Ji-Soo Keum, Sang-Il Oh and Kyung-Nam Kim; Funding acquisition, Jie-Hyun Kim; Investigation, Jie-Hyun Kim; Methodology, Jie-Hyun Kim; Project administration, Jie-Hyun Kim; Resources, Yuna Kim, Jie-Hyun Kim, Jae-Young Chun, Young-Hoon Yoon and Hyojin Park; Software, Ji-Soo Keum, Sang-Il Oh and Kyung-Nam Kim; Supervision, Jie-Hyun Kim and Jae-Young Chun; Validation, Ji-Soo Keum; Visualization, Yuna Kim and Ji-Soo Keum; Writing – original draft, Yuna Kim and Ji-Soo Keum; Writing – review & editing, Yuna Kim, Ji-Soo Keum and Jie-Hyun Kim. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by a grant from the Korean Gastrointestinal Endoscopy Research Foundation (2022 Investigation Grant).

Institutional Review Board Statement: This study was approved by the Institutional Review Board of the Gangnam Severance Hospital (approval no. 3-2022-0309).

Informed Consent Statement: The requirement for informed consent was waived due to the retrospective study design.

Data Availability Statement: The original contributions presented in this study are included in the article/supplementary material. Further inquiries can be directed to the corresponding authors.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial intelligence
CADe	Computer-Aided Detection
CADx	Computer-Aided Diagnosis
CNNs	Convolutional neural networks
VGG	Visual Geometry Group

 doi:10.20944/preprints202503.0938.v1

TP	True positive
TN	True negative
FN	False negative
FP	False positive
PPV	Positive predictive value
ROC	Receiver operating characteristics

References

1. Tavanapong, W.; Oh, J.; Riegler, M.A.; Khaleel, M.; Mittal, B.; de Groen, P.C. Artificial Intelligence for Colonoscopy: Past, Present, and Future. *IEEE J Biomed Health Inform* **2022**, *26*, 3950-3965, doi:10.1109/jbhi.2022.3160098.
2. Jha, D.; Ali, S.; Tomar, N.K.; Johansen, H.D.; Johansen, D.; Rittscher, J.; Riegler, M.A.; Halvorsen, P. Real-Time Polyp Detection, Localization and Segmentation in Colonoscopy Using Deep Learning. *IEEE Access* **2021**, *9*, 40496-40510, doi:10.1109/ACCESS.2021.3063716.
3. Wan, J.; Chen, B.; Yu, Y. Polyp Detection from Colorectum Images by Using Attentive YOLOv5. *Diagnostics (Basel)* **2021**, *11*, doi:10.3390/diagnostics11122264.
4. Biffi, C.; Salvagnini, P.; Dinh, N.N.; Hassan, C.; Sharma, P.; Antonelli, G.; Awadie, H.; Bernhofer, S.; Carballal, S.; Dinis-Ribeiro, M.; et al. A novel AI device for real-time optical characterization of colorectal polyps. *npj Digital Medicine* **2022**, *5*, 84, doi:10.1038/s41746-022-00633-6.
5. Ali, S. Where do we stand in AI for endoscopic image analysis? Deciphering gaps and future directions. *npj Digital Medicine* **2022**, *5*, 184, doi:10.1038/s41746-022-00733-3.
6. Nogueira-Rodríguez, A.; Reboiro-Jato, M.; Glez-Peña, D.; López-Fernández, H. Performance of Convolutional Neural Networks for Polyp Localization on Public Colonoscopy Image Datasets. *Diagnostics (Basel)* **2022**, *12*, doi:10.3390/diagnostics12040898.
7. Nogueira-Rodriguez, A.; Glez-Pena, D.; Reboiro-Jato, M.; Lopez-Fernandez, H. Negative Samples for Improving Object Detection-A Case Study in AI-Assisted Colonoscopy for Polyp Detection. *Diagnostics (Basel)* **2023**, *13*, doi:10.3390/diagnostics13050966.
8. Silva, J.; Histace, A.; Romain, O.; Dray, X.; Granado, B. Toward embedded detection of polyps in WCE images for early diagnosis of colorectal cancer. *Int J Comput Assist Radiol Surg* **2014**, *9*, 283-293, doi:10.1007/s11548-013-0926-3.
9. Bernal, J.; Tajkbaksh, N.; Sanchez, F.J.; Matuszewski, B.J.; Hao, C.; Lequan, Y.; Angermann, Q.; Romain, O.; Rustad, B.; Balasingham, I.; et al. Comparative Validation of Polyp Detection Methods in Video Colonoscopy: Results From the MICCAI 2015 Endoscopic Vision Challenge. *IEEE Trans Med Imaging* **2017**, *36*, 1231-1249, doi:10.1109/tmi.2017.2664042.
10. Jha, D.; Smedsrud, P.H.; Riegler, M.A.; Halvorsen, P.; de Lange, T.; Johansen, D.; Johansen, H.D. Kvasir-SEG: A Segmented Polyp Dataset. Cham, 2020; pp. 451-462.
11. Ma, Y.; Chen, X.; Cheng, K.; Li, Y.; Sun, B. LDPolypVideo Benchmark: A Large-Scale Colonoscopy Video Dataset of Diverse Polyps. Cham, 2021; pp. 387-396.
12. Li, K.; Fathan, M.I.; Patel, K.; Zhang, T.; Zhong, C.; Bansal, A.; Rastogi, A.; Wang, J.S.; Wang, G. Colonoscopy polyp detection and classification: Dataset creation and comparative evaluations. *PLoS One* **2021**, *16*, e0255809, doi:10.1371/journal.pone.0255809.
13. Ali, S.; Jha, D.; Ghatwary, N.; Realdon, S.; Cannizzaro, R.; Salem, O.E.; Lamarque, D.; Daul, C.; Riegler, M.A.; Anonsen, K.V.; et al. A multi-centre polyp detection and segmentation dataset for generalisability assessment. *Sci Data* **2023**, *10*, 75, doi:10.1038/s41597-023-01981-y.
14. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* **2014**.
15. Kim, J.H.; Oh, S.I.; Han, S.Y.; Keum, J.S.; Kim, K.N.; Chun, J.Y.; Youn, Y.H.; Park, H. An Optimal Artificial Intelligence System for Real-Time Endoscopic Prediction of Invasion Depth in Early Gastric Cancer. *Cancers (Basel)* **2022**, *14*, doi:10.3390/cancers14236000.
16. Canbek, G.; Sagiroglu, S.; Temizel, T.T.; Baykal, N. Binary classification performance measures/metrics: A comprehensive visualized roadmap to gain new insights. In Proceedings of the 2017 International Conference on Computer Science and Engineering (UBMK), 5-8 Oct. 2017, 2017; pp. 821-826.

 doi:10.20944/preprints202503.0938.v1

17. Hicks, S.A.; Strümke, I.; Thambawita, V.; Hammou, M.; Riegler, M.A.; Halvorsen, P.; Parasa, S. On evaluation metrics for medical applications of artificial intelligence. *Scientific Reports* **2022**, *12*, 5979, doi:10.1038/s41598-022-09954-8.
18. Sinagra, E.; Badalamenti, M.; Maida, M.; Spadaccini, M.; Maselli, R.; Rossi, F.; Conoscenti, G.; Raimondo, D.; Pallio, S.; Repici, A.; et al. Use of artificial intelligence in improving adenoma detection rate during colonoscopy: Might both endoscopists and pathologists be further helped. *World J Gastroenterol* **2020**, *26*, 5911-5918, doi:10.3748/wjg.v26.i39.5911.
19. Biscaglia, G.; Cocomazzi, F.; Gentile, M.; Loconte, I.; Mileti, A.; Paolillo, R.; Marra, A.; Castellana, S.; Mazza, T.; Di Leo, A.; et al. Real-time, computer-aided, detection-assisted colonoscopy eliminates differences in adenoma detection rate between trainee and experienced endoscopists. *Endosc Int Open* **2022**, *10*, E616-e621, doi:10.1055/a-1783-9678.
20. Wang, K.W.; Dong, M. Potential applications of artificial intelligence in colorectal polyps and cancer: Recent advances and prospects. *World J Gastroenterol* **2020**, *26*, 5090-5100, doi:10.3748/wjg.v26.i34.5090.
21. Li, M.D.; Huang, Z.R.; Shan, Q.Y.; Chen, S.L.; Zhang, N.; Hu, H.T.; Wang, W. Performance and comparison of artificial intelligence and human experts in the detection and classification of colonic polyps. *BMC Gastroenterol* **2022**, *22*, 517, doi:10.1186/s12876-022-02605-2.
22. Wang, W.; Tian, J.; Zhang, C.; Luo, Y.; Wang, X.; Li, J. An improved deep learning approach and its applications on colonic polyp images detection. *BMC Medical Imaging* **2020**, *20*, 83, doi:10.1186/s12880-020-00482-3.
23. Lee, J.Y.; Jeong, J.; Song, E.M.; Ha, C.; Lee, H.J.; Koo, J.E.; Yang, D.-H.; Kim, N.; Byeon, J.-S. Real-time detection of colon polyps during colonoscopy using deep learning: systematic validation with four independent datasets. *Scientific Reports* **2020**, *10*, 8379, doi:10.1038/s41598-020-65387-1.