scientific reports

OPEN

Vision language models versus machine learning models performance on polyp detection and classification in colonoscopy images

Mohammad Amin Khalafi[1,11], Seyed Amir Ahmad Safavi-Naini[1,2,3,11], Ameneh Salehi[1], Nariman Naderi[1], Dorsa Alijanzadeh[1], Pardis Ketabi Moghadam[1], Kaveh Kavousi[4], Negar Golestani[2], Shabnam Shahrokh[1], Soltanali Fallah[5], Jamil S. Samaan[6], Nicholas P. Tatonetti[7,8,9], Nicholas Hoerter[10], Girish Nadkarni[2,3], Hamid Asadzadeh Aghdaei[1]✉ & Ali Soroush[2,3,10]✉

Medical image analysis is central to clinical decision-making, and recent advances in vision–language models (VLMs) have introduced promising capabilities for jointly processing visual and textual data. This study evaluates zero-shot VLMs against convolutional neural networks (CNNs) and classical machine learning (CML) models for polyp detection (CADe) and classification (CADx) using 2,258 colonoscopy images from 428 patients with histopathological labels. We benchmarked 15 approaches including ResNet50, five CMLs (random forest, support vector machine, logistic regression, decision tree, Gaussian naive Bayes), two contrastive vision–language encoders (CLIP, BiomedCLIP), and seven frontier VLMs (GPT-4, GPT-4.1, GPT-4.1-mini, Gemma-3-27b, Qwen-2.5-vl-72b, Gemini-1.5-Pro, Claude-3-Opus). For polyp detection, the highest-performing VLMs (GPT-4.1 F1: 91.98%, GPT-4.1-mini F1: 91.16%) matched CNN performance (ResNet50 F1: 91.35%), though substantial variability existed across VLMs (F1 range: 19.37% to 91.98%). For classification, CNNs substantially outperformed VLMs: ResNet50 achieved weighted F1 of 74.94% versus 55.07% for GPT-4.1-mini, with performance gaps widening dramatically for rare polyp subtypes where VLMs often achieved 0% F1. External validation on 75 images showed that while ResNet50 performance declined substantially, some VLMs demonstrated more stable cross-institutional performance. These findings establish a task-dependent performance hierarchy where VLMs match CNNs for detection but remain limited for classification, suggesting distinct clinical roles for each approach.

Keywords Vision language models, Gastroenterology, Computer aided detection, Colonoscopy, Computer aided diagnosis

Abbreviations

CML	Classical machine learning

[1]Research Institute for Gastroenterology and Liver Diseases, Shahid Beheshti University of Medical Sciences, Tehran, Iran. [2]Division of Data-Driven and Digital Health (D3M), The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [3]The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [4]Laboratory of Complex Biological Systems and Bioinformatics (CBB), Department of Bioinformatics, Institute of Biochemistry and Biophysics (IBB), University of Tehran, Tehran, Iran. [5]Department of GI Diseases, Tehran Milad Hospital, Tehran, Iran. [6]Karsh Division of Gastroenterology and Hepatology, Cedars-Sinai Medical Center, Los Angeles, CA, USA. [7]Department of Computational Biomedicine, Cedars-Sinai Medical Center, West Hollywood, CA, USA. [8]Cedars-Sinai Cancer, Cedars-Sinai Medical Center, 8700 Beverly Blvd, Los Angeles, CA, USA. [9]Department of Biomedical Informatics, Columbia University, New York, NY, USA. [10]Henry D. Janowitz Division of Gastroenterology, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [11]Mohammad Amin Khalafi and Seyed Amir Ahmad Safavi-Naini contributed equally to this work. ✉email: hamid.assadzadeh@gmail.com; Ali.Soroush@mountsinai.org

VLM	Vision language model
cVL	contrastive vision-language encoders
AC	Adenocarcinoma
TA	Tubular adenoma
TVA	Tubulovillous adenoma
VA	Villous adenoma
HP	Hyperplastic polyp
IP	Inflammatory polyp

Colonoscopy remains the gold standard for colorectal cancer screening, yet its effectiveness is fundamentally limited by operator-dependent variability in polyp detection and characterization[1,2]. These limitations have motivated the development of artificial intelligence systems to assist clinicians during colonoscopy: computer-aided detection (CADe) systems that help identify polyps in real time, and computer-aided diagnosis (CADx) systems that suggest the likely histological type based on visual appearance. Such tools aim to reduce missed lesions, improve diagnostic accuracy, and help standardize the quality of colonoscopy across different practice settings[3].

Deep learning approaches have transformed colorectal cancer detection and diagnosis across multiple clinical applications. Convolutional neural networks (CNN) pretrained on large image datasets and fine-tuned on medical images have demonstrated robust performance not only for polyp detection and classification during colonoscopy, but also for histopathological subtyping, prognostic prediction from tissue samples, and treatment response assessment. Recent architectures including YOLO variants, enhanced U-Net models, and transformer-based approaches have achieved detection sensitivities exceeding 90% and real-time processing capabilities suitable for clinical deployment[4–9] (Table 1). However, the CNN paradigm imposes significant development constraints. Each new model requires extensive labeled training data specific to the target population and imaging equipment, iterative optimization of network architectures and hyperparameters, and validation across multiple institutions to ensure generalizability. These requirements make CNN development resource-intensive and create barriers to rapid adaptation as clinical needs or imaging technology evolve.

Advances in vision-language models (VLM) suggest an alternative approach that addresses data and development barriers. Contrastive Language-Image Pre-training (CLIP) demonstrated that joint training of visual and language encoders on large-scale image-text pairs enables zero-shot task performance through natural language prompts alone[10]. Unlike CNNs that require fine-tuning on labeled medical images to adapt pretrained features to specific tasks, CLIP-based models can be deployed directly through prompt specification. BiomedCLIP extended this framework to the biomedical domain through pretraining on 15 million figure-caption pairs from PubMed Central[11], improving medical imaging performance while maintaining zero-shot deployment. More recently, large VLM such asGPT-4[12], Claude-3-Opus[13], and Gemini-1.5-Pro[14] have integrated sophisticated visual encoders with transformer-based language models, enabling complex reasoning about medical images without any task-specific fine-tuning (Table 2)[15–21]. These models represent a fundamentally different deployment paradigm: rather than adapting model weights to each clinical task, the same pretrained model is applied across diverse applications through natural language instructions.

The potential advantages of zero-shot VLM for CADe and CADx are substantial. Eliminating the fine-tuning step removes the need for institution-specific labeled datasets and model optimization. Prompt-based interaction allows flexible task specification without retraining. Pretraining on billions of diverse images may confer robustness to the distribution shifts that degrade fine-tuned CNN performance across institutions. However, these theoretical advantages remain unvalidated for colonoscopy applications. Whether zero-shot VLMs can match the detection sensitivity of fine-tuned CNNs, how they perform across histological classification tasks of varying difficulty, whether they generalize better to external datasets, and how sensitive they are to prompt design are empirical questions with direct implications for clinical deployment strategies.

We systematically evaluated 15 computational approaches spanning classical machine learning (CML), CNN, contrastive vision-language encoders, and state-of-the-art VLMs for frame-level polyp detection and histological classification. Using 2,258 colonoscopy images with pathological confirmation and external validation on 75 images from an independent institution, we compared zero-shot VLM performance against

First Author, Year	VLM \Model	Major	Modality	Performance/Contribution
Pecal, 2021[4]	YOLOv3 + CSPNet, SiLU	Gastroenterology, polyp detection	Colonoscopy	Improved YOLOv3/YOLOv4 with higher precision/recall; validated on large datasets, enhancing clinical usability.
Karaman, 2023b[5]	YOLOv5 + ABC optimization	Gastroenterology, polyp detection	Colonoscopy	ABC-tuned hyperparameters and activations; outperformed baseline YOLOv5 in accuracy and speed
Karaman, 2023a[6]	Scaled-YOLOv4 + ABC	Gastroenterology, polyp detection	Colonoscopy	First systematic YOLO optimization; +3% mAP and + 2% F1 across multiple variants.
Pecal and Karaboga, 2021[7]	YOLOv4 + CSPNet, Mish, ensemble	Gastroenterology, polyp detection	Colonoscopy	State-of-the-art detection with precision 96%, recall 97%, F1 96%; real-time applicability.
Narasimha Raju, 2025[9]	Hybrid CNN (ResNet-50, DenseNet-201, VGG-16) + Transformer + Multi-class SVM + Grad-CAM	CRC (multi-class lesion detection)	Colonoscopy	Achieved 98% accuracy, F1 = 0.98, precision = 97%, recall = 99%. Addressed class imbalance, interpretability, and spatial complexity with explainable heatmaps; sets new benchmark for clinically interpretable AI-assisted colonoscopy

Table 1. Overview of studies assessing the performance of deep learning models in medical imaging.

First Author, Year	VLM \Model	Major	Modality	Performance/Contribution
Pilia, 2024; and Hardin, 2024[15]	GPT-4	Dermatology	Image/ Scenario Prompt /Image + Scenario Prompt	GPT-4 V accuracy: image-only: 54%/ text-only scenarios: 89%/ both image + scenario: 89%
Laohawetwanti, 2024[16]	custom GPT-4	Histopathology	Colorectal polyp photomicrographs	GPT-4 accuracy: 16% for non-specific changes / 36% for tubular adenomas Sensitivity: 74% for adenoma detection specificity: 36% for adenoma detection
Chen, 2023[17]	GPT-4 V	Internal medicine	COVID-19 lung X-ray	GPT4-V accuracy: ranged 72% to 85% based on different prompts.
Han, 2023[18]	GPT-4	General Medicine	Clinical cases from the JAMA Clinical Challenge and the NEJM Image Challenge	GPT-4 V accuracy: 73.3% for JAMA and 88.7% for NEJM
Xu, 2024[19]	GPT-4	ophthalmology	various ocular imaging modalities	Examination Identification :95.6% Lesion Identification:25.6% Diagnosis Capacity:16.1% Decision Support:24%
Yang, 2023[20]	GPT-4	General Medicine	USMLE with Image	For questions with images: 86.2%, 73.1%, and 62.0% on USMLE, DRQCE, and AMBOSS. For questions with image, GPT-4 achieved an accuracy of 84.2%, 85.7%, 88.9% in Step1, Step2CK, and Step3 of USMLE questions
Jin, 2024[21]	GPT-4	General Medicine	Clinical cases from NEJM Image Challenges + scenario prompt	GPT-4 accuracy: 81.6%, which outperformed physicians and medical students.

Table 2. Overview of studies assessing the performance of vision Language models in medical imaging.

fine-tuned CNNs and classical methods across both binary detection and multi-class classification tasks. We further investigated prompt engineering strategies, few-shot learning, computational requirements, and model interpretability to assess practical deployment considerations. Our results establish performance benchmarks across model families, reveal task-dependent capabilities and limitations, and provide evidence-based guidance for selecting appropriate approaches based on clinical requirements and available resources.

Methods

Ethical consideration

This study received ethical approval from the Institutional Review Board at the Research Ethics Committees of the Research Institute for Gastroenterology & Liver Diseases at Shahid Beheshti University of Medical Sciences (IR.SBMU.RIGLD.REC.1401.043). In accordance with the principles outlined in the Helsinki Declaration, patient confidentiality and welfare was maintained throughout the study. All procedures involving patient data and images were conducted using standardized protocols to safeguard patient privacy, with measures in place to anonymize data and prevent identification. Explicit informed consent was obtained from all participants, affirming their voluntary participation in the study.

The external dataset used in this study was anonymized and obtained under a signed data agreement. The dataset provider had secured prior ethical approval for its collection and use, and is registered in National Registry of Biobanks (B.0000140) and ISCIII Biomodels and Biobanks Platform (PT23/00013).

Experimental framework

This investigation followed a retrospective, comparative methodological design to evaluate multiple artificial intelligence approaches for colonoscopy image analysis. We adhered to Consolidated reporting guidelines for prognostic and diagnostic machine learning modeling studies[22] and the transparent reporting of a multivariable prediction model for Individual prognosis or diagnosis (TRIPOD-AI)[23] guidelines for model development and results reporting, ensuring methodological transparency and reproducibility. We structured our investigation as a three-phase experimental program designed to systematically evaluate model performance:

1. **Parameter Optimization Phase**: We systematically identified optimal hyperparameters for each model architecture through comprehensive grid search methodologies, establishing optimized configurations for subsequent performance evaluation.
2. **Detection Evaluation Phase**: We conducted comparative assessments of model performance in identifying polyp presence (CADe functionality), utilizing standardized metrics, including F1 scores and area under the receiver operating characteristic curve (AUROC).
3. **Classification Analysis Phase**: We performed systematic evaluation of model efficacy in correctly classifying polyp pathology types (CADx functionality) across six distinct histological categories, employing weighted evaluation metrics to account for class distribution.

This structured approach enabled comprehensive, controlled comparison across diverse computational methodologies while maintaining consistent evaluation standards.

Dataset - Characteristics

Patient population and data collection

We examined colonoscopy data collected between December 2022 and April 2023 at Taleghani Hospital's gastroenterology clinic and Behbood clinic. The study population comprised 428 patients (mean age: 53 ± 14 years; 48.6% male) who underwent colonoscopy for primary colorectal cancer screening, post-polypectomy surveillance, evaluation following positive fecal immunochemical tests, or investigation of gastrointestinal symptoms.

All procedures were performed by gastroenterologists with extensive experience (> 2,000 screening colonoscopies conducted). The endoscopists assessed bowel preparation quality using the validated Boston Bowel Preparation Scale and confirmed cecal intubation through identification of the ileocecal valve and appendix orifice.

Image collection and histopathological assessment

We compiled a comprehensive image dataset consisting of 1,129 colon polyp images and 1,129 randomly selected normal colon images (from an original pool of 6,046) to address class imbalance. The initial classification was derived from procedure pathology reports, followed by an expert review of stored images by an experienced gastroenterologist (PKM) who assigned final labels.

Tissue samples underwent standard histopathological processing, including formalin fixation, paraffin embedding, sectioning (4–5 microns thick), and hematoxylin-eosin staining. Histological classification followed established criteria, with assessment of cellular atypia, glandular architecture, and dysplasia degree[24]. Our final dataset comprised 2,258 images from 428 patients, including tubular adenoma ($n = 771$), hyperplastic polyp ($n = 138$), adenocarcinoma ($n = 79$), tubulovillous adenoma ($n = 59$), inflammatory polyp ($n = 45$), villous adenoma ($n = 36$), and normal colon ($n = 1{,}129$). Complete dataset characteristics are provided in Table 3.

External dataset for validation

Sample images and anonymized patient data used in this study were obtained from the PICCOLO database of the Basque Biobank (www.biobancovasco.bioef.eus*)*, which is registered in the National Registry of Biobanks (B.0000140) and integrated into the ISCIII Biomodels and Biobanks Platform (PT23/00013). This dataset contains 3433 images from clinical colonoscopy videos, including white light and narrow band imaging (NBI) images, from colonoscopy procedures in human patients. It includes 76 different lesions from 48 patients. We selected a total of 75 images, comprising 9 adenocarcinomas, 50 adenomatous polyps, and 16 hyperplastic polyps from white light images.

Since the external dataset contains no normal images and only three distinct polyp classes, we adapted our internal dataset by selecting and organizing it in the same way, allowing for a consistent comparison between internal and external datasets.

Image preprocessing and data augmentation

We implemented a comprehensive preprocessing pipeline to optimize image quality and enhance model training. All images underwent uniform resizing to 300 × 300 pixels, followed by normalization to standardize pixel value distribution. To enhance model robustness and generalizability, we applied a systematic augmentation protocol incorporating horizontal and vertical mirroring to diversify polyp orientation representation, brightness variations to simulate diverse lighting conditions, Gaussian blur application to replicate optical aberrations, additive Gaussian noise to build resilience against image artifacts, and linear contrast adjustments to enhance structural differentiation. This augmentation strategy yielded a four-fold expansion of the effective training dataset, simultaneously enhancing model exposure to diverse image acquisition parameters and reducing overfitting to institution-specific imaging characteristics.

Category	Case (Total)	Train	Test	Control (Total)	Train	Test
Patients (n)	237			191		
Age (mean ± SD)	55 ± 13			50 ± 14		
Male (n, %)	130 (54.8%)			78 (40.8%)		
Images (n)	1232			1025		
Normal (n)		82	22		805	220
Adenocarcinoma (n)		66	13		-	-
Tubular adenoma (n)		650	121		-	-
Tubulovillous adenoma (n)		48	11		-	-
Villous adenoma (n)		30	6		-	-
Hyperplastic polyp (n)		116	22		-	-
Inflammatory polyp (n)		38	7		-	-

Table 3. Characteristics of the dataset at both patient and image levels.

Model development and configuration

Classical machine learning approaches

We implemented five distinct classical machine learning algorithms, each optimized through systematic hyperparameter tuning (Table 4). For the Decision Tree Classifier, we employed a comprehensive grid search across multiple parameters, including criterion ('gini', 'entropy'), max_depth (None, 10, 20, 30), min_samples_split (2, 5, 10), and min_samples_leaf (1, 2, 4). The optimal configuration identified was criterion='entropy', max_depth = 20, min_samples_leaf = 2, and min_samples_split = 2. For the Random Forest Classifier, our hyperparameter optimization encompassed n_estimators (50, 100, 200), max_depth (None, 10, 20, 30), min_samples_split (2, 5, 10), and min_samples_leaf (1, 2, 4). The optimal configuration was determined to be n_estimators = 200, min_samples_leaf = 1, min_samples_split = 10, and random_state = 42. With the Support Vector Machine (SVM), we systematically evaluated parameter combinations including C (0.1, 1, 10), kernel ('linear', 'rbf', 'poly'), and gamma ('scale', 'auto'). The optimal configuration identified was kernel='rbf', C = 10, gamma='scale', probability = True, and random_state = 42. For Logistic Regression, our grid search evaluated C values (0.1, 1, 10) and solver options ('newton-cg', 'lbfgs', 'liblinear', 'sag', 'saga'). The optimal configuration was determined to be C = 0.1, solver='sag', and random_state = 42. The Gaussian Naive Bayes algorithm was implemented with default parameters as it does not feature adjustable hyperparameters.

Convolutional neural network: Resnet 50

We implemented ResNet50 based on its demonstrated superior performance in medical image classification tasks[25]. To optimize performance, we conducted systematic hyperparameter tuning via GridSearchCV, evaluating learning_rate (0.01, 0.1, 1), epochs (5, 10, 15), and batch_size (32, 64). The grid search involved dividing the dataset into training, validation, and testing subsets, training the model on various hyperparameter combinations, and using cross-validation to evaluate performance and prevent overfitting. The optimal configuration was determined to be learning_rate = 0.01, epochs = 15, and batch_size = 32.

Contrastive multimodal encoders

We evaluated two specialized contrastive learning models for our analysis. CLIP represents a general-purpose multimodal model that associates images with corresponding textual descriptions through dual visual and textual encoders trained on 400 million image-text pairs[10]. We implemented the ViT-B/32 variant for zero-shot evaluation in our experimental framework. Additionally, we assessed BiomedCLIP, a domain-specialized adaptation of CLIP that underwent pretraining on PMC-15 M, a dataset comprising 15 million biomedical figure-caption pairs from PubMed Central publications[11]. This biomedical specialization potentially enhances performance for medical imaging applications, making it particularly relevant for our colonoscopy image analysis.

General-Purpose vision Language models

We evaluated seven state-of-the-art VLMs as part of our comprehensive assessment. GPT-4 represents an enhanced iteration of OpenAI's GPT-4 model that integrates advanced visual processing capabilities, enabling interpretation of and response to image inputs[26]. In addition, we assessed the performance of state-of-the-art OpenAI models, namely GPT-4.1 and GPT-4.1-mini. We also included Claude-3-Opus, developed by Anthropic, which builds upon their Claude architecture with enhanced visual question answering capabilities[13]. The fifth and sixth models in our evaluation was Gemini-1.5-Pro, Google's multimodal foundation model designed for versatile tasks including visual comprehension, classification, and content generation across modalities[14] and Gemma-3-27B. The last model in our evaluation was Qwen-2.5-VL-72B. These general-purpose models were evaluated without domain-specific fine-tuning to assess their zero-shot capabilities in medical image analysis.

We utilized the web-based API interface of GPT-4 (*gpt-4-1106-vision-preview;* Accessed: May 2024 via API), GPT-4.1 (*Created Apr 14, 2025;* Accessed: August 2025 via API), GPT4.1-mini (*Created Apr 14, 2025;* Accessed: August 2025 via API), Claude-3-Opus (*claude-3-opus-20240229;* Accessed: May 2024 via API), Qwen-2.5-vl-72B (*Created Feb 1, 2025;* Accessed: August 2025 via API), Gemma-3-27B (*Created Mar 12, 2025;* Accessed: August 2025 via API) and Gemini-1.5-Pro (*gemini-1.5-pro-001;* Accessed: June 2024 via Google interface),.

Model Name	Hyperparameters	Best Hyperparameters	Best Accuracy

Decision Tree	criterion ('gini', 'entropy')max_depth (None, 10, 20, 30)
	min_samples_split (2, 5, 10)min_samples_leaf (1, 2, 4)
		criterion='entropy'
		max_depth = 20min_samples_split = 2
		min_samples_leaf = 2
			0.6609

Random Forest	n_estimators (50, 100, 200)max_depth (None, 10, 20, 30)
	min_samples_split (2, 5, 10)
	min_samples_leaf (1, 2, 4)
		n_estimators = 200
		max_depth = None
		min_samples_leaf = 1min_samples_split = 10random_state = 42
			0.7707

Support Vector Machine	C (0.1, 1, 10)
	kernel ('linear', 'rbf', 'poly')
	gamma ('scale', 'auto')
		C = 10
		kernel='rbf'
		gamma='scale'
			0.7780

Logistic Regression	C (0.1, 1, 10)
	solver ('newton-cg', 'lbfgs', 'liblinear', 'sag', 'saga')
		C = 0.1,
		solver ='sag'
			0.7512

Resnet 50	learning_rate (0.01, 0.1, 1)
	epochs (5, 10, 15)
	batch_size (32, 64)
		learning_rate = 0.01,
		epochs = 15,
		batch_size = 32
			0.8842

Table 4. Hyperparameter Tuning.

Approximately 15% of our test dataset was allocated for the parameter optimization phase, while the remaining 85% was used for the detection and classification evaluation phase. All experiments were conducted with standardized parameters (temperature = 1.0, maximum tokens = 512, tool calls disabled, random seed = 123) to ensure consistent evaluation conditions. We renamed all image file names to avoid any data leakage from the image metadata.

To assess the impact of prompt optimization, we first used the following raw prompt in a chat: "*What is this image?*" accompanied by the image. Subsequently, in the same chat, we asked: "*What is the pathology class of the polyp? Give me only one answer.*" In a separate chat, we then posed this engineered prompt:

> "*As an esteemed gastroenterologist specializing in colonoscopy evaluation, your expertise is crucial in meticulously assessing a provided colonoscopy image. Your task is to discern and characterize any irregularities present across the colonic mucosa, paying close attention to morphology, color variations, and vascularity patterns. Drawing upon your wealth of experience, construct a comprehensive list of potential diagnoses, including but not limited to inflammatory bowel disease, colorectal polyps, diverticulosis, and colorectal cancer. Your discerning analysis and diagnostic acumen will guide subsequent clinical decisions, emphasizing the importance of accurate interpretation and effective communication in delivering optimal patient care.*"

This was followed by the image. Then, in the same chat, we used the prompt:

> "*Analyze the provided image and select one of the following options that accurately describes the patient's diagnosis*:

1. *normal*.
2. *adenocarcinoma*.
3. *adenomatous-tubular polyp*.
4. *adenomatous-tubulovillous polyp*.
5. *adenomatous-villous polyp*.
6. *hyperplastic polyp*.
7. *inflammatory polyp.*"

The optimized prompt was developed through a human-in-the-loop refinement process whereby candidate variations were generated using GPT-4, informed by prompt engineering techniques adapted from validated gastroenterology-specific methodologies[27]. These techniques included contextual embedding (providing task-specific domain context), expert mimicry (emulating clinical specialist reasoning patterns), chain-of-thought reasoning (eliciting stepwise analytical processes), exemplar anchoring (supplying representative clinical scenarios), and constrained output formatting (defining structured response schemas). A domain expert subsequently reviewed and refined the candidate to produce the final optimized prompt.

Exploring Few-Shot injection impact on General-Purpose vision Language models

For few-shot learning, we selected representative images directly from the training dataset to serve as illustrative examples for the model. Specifically, we curated two sets of images with corresponding labels. The first set (1 image for 'no-polyp' and 1 image for 'polyp') focused on distinguishing between polyp and non-polyp cases, providing general guidance on the presence or absence of polyps. The second set (one image for 'normal' and one image for each polyp subtypes) concentrated on specific pathology classes. Each few-shot example consisted of an image paired with a descriptive label, and these were included in the prompt to the model to facilitate accurate and informed predictions on unseen test images. We applied few-shot learning to recently released, state-of-the-art VLMs (GPT-4.1, GPT-4.1-mini, Qwen-2.5-vl-72b and Gemma-3-27b).

Performance evaluation

We developed an approach that converts unstructured text into structured classifications using GPT-4 to facilitate the semi-automated evaluation of textual outputs. The model was configured with a temperature setting of 0 and enabled to generate structured JSON outputs.

The extraction system was designed to categorize VLM responses into predefined labels with explicit handling of uncertain or ambiguous cases. For polyp detection, the system classified responses into: (1) "Human evaluation needed: I am unsure," (2) "Human evaluation needed: More than one diagnosis is selected, or no option is selected," (3) "The unstructured answer selected: No polyp is detected in the image," or (4) "The unstructured answer selected: A polyp is detected in the image." For polyp classification, an additional category was included: (5) "The unstructured answer selected: The polyp type is classified as {polyp_type in polyp_types}."

This structured extraction approach enables automated classification while flagging ambiguous or uncertain cases for human review, ensuring accuracy in the evaluation process. The system processes free-text responses by identifying key diagnostic terminology, matching it to predefined categories, and assigning confidence scores. Responses containing hedging language ("possibly," "might be," "unclear") or multiple conflicting diagnoses were automatically flagged for human review.

To validate this approach, we manually reviewed a random sample of 50 response-extraction pairs. GPT-4 correctly extracted and labeled all 43 unambiguous responses while appropriately flagging 7 cases requiring human evaluation, demonstrating 100% accuracy for clear cases and appropriate conservative handling of ambiguous outputs. All flagged cases were subsequently reviewed by a clinical expert (PKM) to assign final labels.

Statistical analysis

We performed comprehensive statistical analysis using Python (version 3.11.5), employing standardized machine learning evaluation methodologies. We implemented the one-vs-all strategy for multiclass classification scenarios to enable binary performance metrics for each class. We selected the F1 score as our primary evaluation metric due to its balanced consideration of both precision and recall, making it particularly suitable for our dataset where class imbalance was present, especially in the polyp classification tasks where some pathology types had limited representation.

Performance was evaluated using multiple complementary metrics: F1 scores to balance precision and recall considerations; AUROC to assess discriminative capability; confusion matrices to visualize classification patterns and error types; and weighted metrics to account for class imbalance in overall performance assessment. For weighted F1 scores in polyp classification tasks, we calculated values based on the proportion of each polyp type in the test dataset, ensuring that performance metrics appropriately reflected the distribution of classes in clinical settings.

TiLense: importance of tiles for vlm's Zero-Shot polyp detection

This proposed approach seeks to identify and visualize key image tiles in vision-language tasks by assessing the significance of each tile through frequent responses across multiple prediction attempts. In contrast to complex methods, it focuses on a single, dominant answer instead of the original model probability. The procedure involves pinpointing the primary answer, evaluating tile significance, and then generating a heatmap to showcase these important areas. This model-agnostic unsupervised technique elucidates essential regions in VLM classification by juxtaposing tile-based results with a singular base answer after N iterations. By highlighting areas where significant variations occur, it uncovers which sections of an image most influence model predictions, which is beneficial for evaluation and improvement. We refer to this method as "TiLense" due to its capacity to highlight importance across image tiles for zero-shot prediction tasks.

We implemented this tile masking technique to showcase GPT-4.1 and GPT-4's vision capabilities in zero-shot prediction tasks across four scenarios: the presence of a polyp, a polyp in a challenging background, a standard image, and a standard image in a complex background. A systematic sliding window approach masked specific regions of the images (see Fi.g. 3a). The original and masked images were evaluated by GPT-4 using a standardized prompt, with a temperature setting of 1, a maximum token limit of 300, and no specified seed value, with the process repeated five times to create response distributions. The base answers were established through majority voting. The output is represented as a heatmap, where each tile is colored according to its impact on altering the base answer. Since tiles can overlap, we scale each tile from 0 to 1, coloring them from white to red.

Libraries and local computing

For our VLM API calls, we used Python 3.11.5 in combination with the "requests" library, enabling efficient interaction with computational resources. Local experiments for CMLs and ResNet50 training and testing were conducted on a laptop equipped with a Ryzen 7–4800 H CPU and 16 GB of RAM, where we employed the scikit-learn and TensorFlow libraries for model implementation and evaluation.

Results

Model optimization

Our initial experimental phase focused on optimizing model configurations and prompt strategies for VLMs. We observed that domain-specific prompts consistently outperformed simple queries across all VLMs tested. For polyp detection, the smallest improvement was observed with Gemini-1.5-Pro (F1: from 0.715 to 0.731; +2.2%), while the largest gain was achieved by Qwen-2.5-vl-72b (F1: from 0.531 to 0.802; +51.0%). For polyp classification, the minimum improvement was seen in Claude-3-Opus (weighted F1: from 0.112 to 0.147; +31.2%), whereas the maximum improvement occurred with Qwen-2.5-vl-72b (weighted F1: from 0.008 to 0.502; +6175.0%). Table 5 provides a detailed comparison of performance improvements across prompting strategies, which formed the foundation for our subsequent analyses and mention prompt engineering techniques that we used. **Supplementary Figures S1** and S2 present the confusion matrices of answers for polyp detection and classification, respectively.

Polyp detection performance (CADe)

Polyp detection performance established a clear hierarchical distribution across models, as demonstrated by confusion matrices (Fig. 1) and F1 scores (Table 6). GPT-4.1 achieved the highest performance (F1: 91.98%), closely followed by ResNet50 (F1: 91.35%) and GPT-4.1-mini (F1: 91.16%), demonstrating that latest-generation VLMs can match task-specific CNNs for binary detection. BiomedCLIP demonstrated strong results (F1: 88.68%), outperforming general CLIP (F1: 68.39%) by more than 20%. Traditional machine learning and earlier VLMs formed the next tier: Random Forest and GPT-4 (both F1: 81.02%), SVM (F1: 77.92%), and Logistic Regression (F1: 72.80%). Moderate capability was observed for Decision Tree (F1: 68.10%), Qwen-2.5-vl-72b (F1: 68.59%), Gemma-3-27b (F1: 69.29%), and Claude-3-Opus (F1: 66.40%). The lowest detection capability was exhibited by Gemini-1.5-Pro (F1: 19.37%) and Gaussian Naive Bayes (F1: 10.22%). Confusion matrices for all models are presented in Fig. 1. AUROC analysis (Fig. 2) reinforced these findings, with top performers achieving values above 0.95.

We applied the TiLense tile-based importance mapping method to elucidate model decision-making processes for GPT-4.1 and GPT-4. Figure 3 presents attention heatmaps across four diagnostically relevant scenarios: normal mucosa (**3b**), standard polyp (**3c**), poorly prepared normal mucosa (**3d**), and subtle polyp (**3e**). GPT-4.1 demonstrated clinically appropriate attention allocation, with high-importance tiles accurately localizing polyp regions in clear cases (**3c**) and maintaining focus on pathologically relevant features across

	Prompt Engineering Technique	GPT-4	Claude-3-Opus	Gemini-1.5-Pro	GPT-4.1	GPT-4.1mini	Qwen-2.5-vl-72b	Gemma-3-27b
		F1 score (change)	F1 score (change)	F1 score (change)	F1 score (change)	F1 score (change)	F1 score (change)	F1 score (change)
Polyp Detection	Raw Prompt[a]	0.636 (ref)	0.266 (ref)	0.715 (ref)	0.915 (ref)	0.915 (ref)	0.531 (ref)	0.652 (ref)
Polyp Detection	Contextual Embedding, Expert Mimicry, Chain of Thought, Anchoring with Examples[b]	0.748 (+ 17.6%)	0.458 (+ 72.2%)	0.731 (+ 2.2%)	0.935 (+ 2.2%)	0.956 (+ 4.5%)	0.802 (+ 51.0%)	0.798 (+ 22.4%)
Polyp Classification	Raw Prompt[c]	0.126 (ref)	0.112 (ref)	0.0 (ref)	0.156 (ref)	0.169 (ref)	0.008 (ref)	0.190 (ref)
Polyp Classification	Constrained Output[d]	0.548 (+ 434.9%)	0.147 (+ 31.2%)	0.437 (NA)	0.594 (+ 280.7%)	0.711 (+ 320.7%)	0.502 (+ 6175.0%	0.350 (+ 84.2%)

Table 5. Impact of prompt engineering on vision Language model performance. a: "What is this image?" b: "As an esteemed gastroenterologist specializing in colonoscopy evaluation, your expertise is crucial in meticulously assessing a provided colonoscopy image. Your task is to discern and characterize any irregularities present across the colonic mucosa, paying close attention to morphology, color variations, and vascularity patterns. Drawing upon your wealth of experience, construct a comprehensive list of potential diagnoses, including but not limited to inflammatory bowel disease, colorectal polyps, diverticulosis, and colorectal cancer. Your discerning analysis and diagnostic acumen will guide subsequent clinical decisions, emphasizing the importance of accurate interpretation and effective communication in delivering optimal patient care." c: "What is the pathology class of the polyp? Give me only one answer." d: "Analyze the provided image and select one of the following options that accurately describes the patient's diagnosis: \nnormal \nadenocarcinoma \n adenomatous-tubular polyp \n adenomatous-tubulovillous polyp \n adenomatous-villous polyp \n hyperplastic polyp \n inflammatory polyp."

varying image quality conditions. In contrast, GPT-4 exhibited attention misallocation in challenging scenarios, incorrectly prioritizing artifacts in poorly prepared images (**3d**) and displaying dispersed attention patterns for subtle lesions (**3e**), revealing susceptibility to image quality degradation and low-contrast pathology. These attention pattern differences align with the models' respective classification accuracies, suggesting that GPT-4.1's performance gains reflect improved capacity to focus on clinically meaningful anatomical features rather than confounding visual elements.

Polyp classification performance (CADx)

Classification performance revealed a different hierarchy than detection, with CNNs substantially outperforming VLMs for fine-grained histological discrimination (Fig. 4). ResNet50 achieved the highest weighted F1 (74.94%), establishing a 20-percentage-point advantage over the best VLMs: GPT-4.1-mini (55.07%) and GPT-4.1 (54.74%). SVM was the only other model exceeding 55% (55.63%). Mid-tier performers included Random Forest (43.67%), Qwen-2.5-vl-72b (42.13%), GPT-4 (41.18%), Logistic Regression (40.32%), and Decision Tree (40.42%). Earlier VLMs and contrastive encoders showed weaker performance: Gemma-3-27b (35.50%), BiomedCLIP (27.74%), Claude-3-Opus (25.54%), Gemini-1.5-Pro (6.17%), and CLIP (1.69%). Notably, BiomedCLIP's strong detection (88.68%) did not translate to classification (27.74%), suggesting zero-shot classification of subtle histological variants is substantially more challenging. Table 6 presents overall weighted F1 scores, while **Supplementary Table**S1 details performance by polyp type.

Tubular adenoma (TA) images (650 training, 121 test) achieved the most consistent classification performance across models. The best results were obtained by ResNet50 (F1: 0.85), followed by Support Vector Machine (F1: 0.68) and Random Forest (F1: 0.64). Among VLMs, GPT-4 (F1: 0.58) outperformed Claude-3-Opus (F1: 0.33). However, other recent VLMs such as Gemma-3-27B (F1: 0.48) and Qwen-2.5-VL-72B (F1: 0.57) showed weaker performance. Notably, the latest multimodal models, GPT-4.1 (F1: 0.71) and GPT-4.1-mini (F1: 0.73), narrowed the gap with CNN and CML methods, underscoring rapid progress in VLM-based polyp subtype recognition.

Adenocarcinoma (AC) images (66 training, 13 test) were best classified by GPT-4.1-mini (F1: 0.69), closely followed by ResNet50 (F1: 0.67); GPT-4.1 (F1: 0.61) trailed both. Among other models, BiomedCLIP (F1: 0.56) and SVM (F1: 0.45) performed reasonably, while tree-based methods were low (Decision Tree: 0.06; Random Forest: 0.00). Other VLMs were modest: GPT-4 (F1: 0.30), Qwen-2.5-VL-72B (F1: 0.25), Gemma-3-27B (F1: 0.24), Claude-3-Opus (F1: 0.19), Gemini-1.5-Pro (F1: 0.00).

Hyperplastic polyp (HP) images (116 training, 22 test) presented a challenging classification task. Among CML methods, SVM (F1: 0.31) and Decision Tree (F1: 0.22) outperformed Random Forest (F1: 0.08), Logistic Regression (F1: 0.07), and Gaussian Naive Bayes (F1: 0.07). The CNN ResNet50 achieved the highest overall performance with an F1 of 0.49, highlighting the strength of deep learning for this subtype. VLMs generally performed poorly: GPT-4 and GPT-4.1-mini (F1: 0.00), Gemini-1.5-Pro (F1: 0.00), while GPT-4.1 (F1: 0.14), Claude-3-Opus (F1: 0.14), Qwen-2.5-vl-72b (F1: 0.09), and Gemma-3-27b (F1: 0.05) performed slightly better. Among contrastive VLMs, BiomedCLIP (F1: 0.21) outperformed CLIP (F1: 0.04) but still lagged behind CNN and CML models.

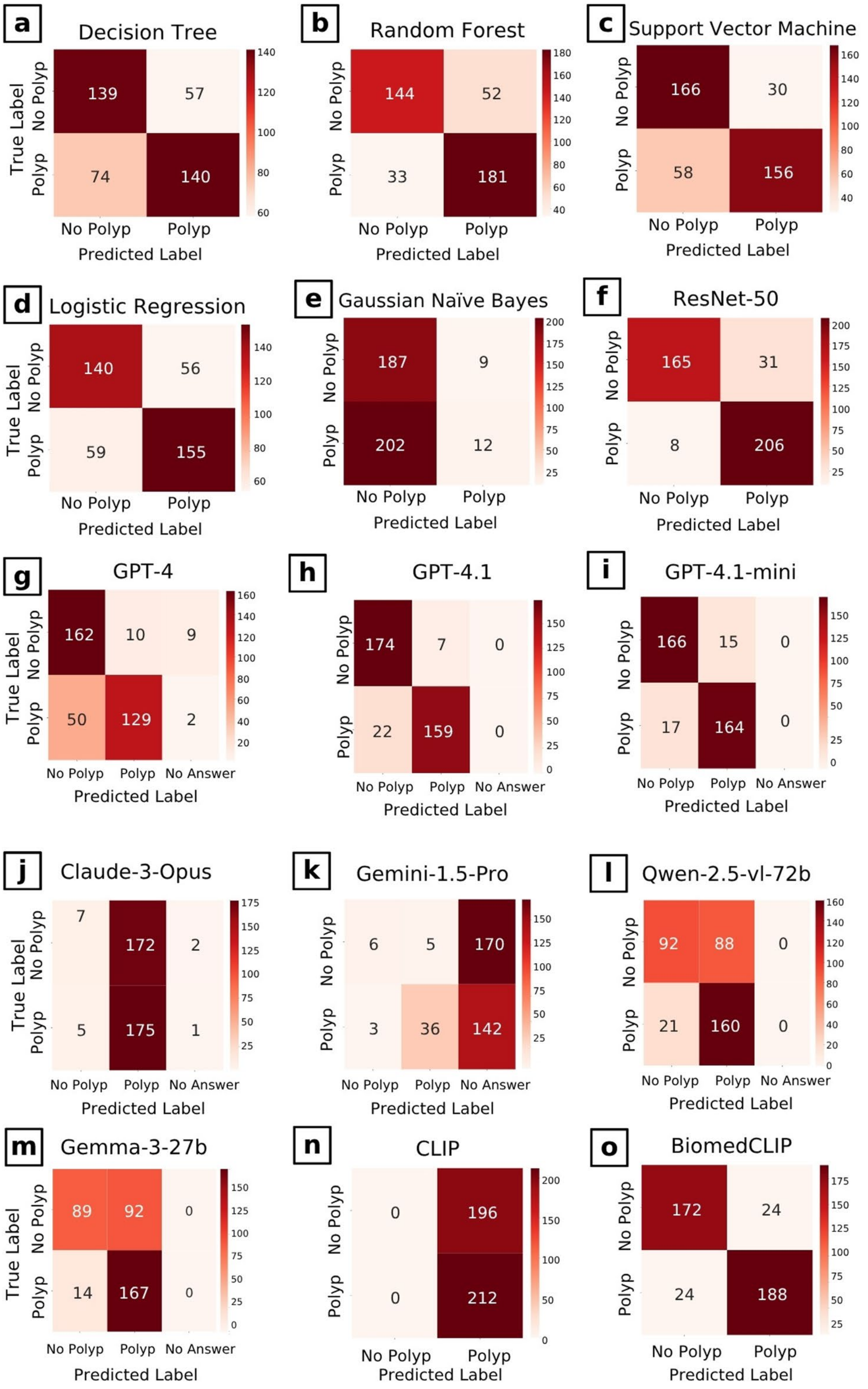

Fig. 1. Polyp detection performance across machine learning and vision language models. Confusion matrices depicting polyp detection performance across various models in test set (internal validation): classical machine learning algorithms—Decision Tree (**a**), Random Forest (**b**), Support Vector Machine (**c**), Logistic Regression (**d**), Gaussian Naive Bayes (**e**); convolutional neural network—ResNet-50 (**f**); vision-language models—GPT-4 (**g**), GPT-4.1 (**h**); GPT-4.1-mini (**i**), Claude-3-Opus (**j**), Gemini-1.5-Pro (**k**), Qwen-2.5-vl-72b (**l**), Gemma-3-27b (**m**); and contrastive vision-language encoders—CLIP (**n**), BiomedCLIP (**o**). Each matrix illustrates model predictions relative to ground-truth labels.

Model Family	Model	Polyp Detection	Polyp Classification	AC (*N* = 79)	TA (*N* = 771)	TVA (*N* = 59)	VA (*N* = 36)	HP (*N* = 138)	IP (*N* = 45)
		F1	Weighted F1	F1	F1	F1	F1	F1	F1
CML	Decision tree	0.681	0.4042	0.06	0.53	0.27	0.00	0.22	0.14
CML	Random forest	0.8102	0.4367	0.00	0.64	0.00	0.00	0.08	0.00
CML	Support vector machine	0.7792	0.5563	0.45	0.68	0.25	0.00	0.31	0.36
CML	Logistic regression	0.728	0.4032	0.10	0.56	0.00	0.00	0.07	0.20
CML	Gaussian naive bayes	0.1022	0.0764	0.08	0.09	0.00	0.00	0.07	0.00
CNN	ResNet50	0.9135	**0.7494**	0.67	**0.85**	**0.55**	**0.25**	**0.49**	**0.71**
VLM	GPT-4	0.8102	0.4118	0.30	0.58	0.00	0.00	0.00	0.00
VLM	Claude-3-Opus	0.664	0.2554	0.19	0.33	0.06	0.04	0.14	0.00
VLM	Gemini-1.5-Pro	0.1937	0.0617	0.00	0.09	0.00	0.00	0.00	0.00
VLM	GPT-4.1	0.9198	0.5474	0.61	0.71	0.00	0.20	0.14	0.08
VLM	GPT-4.1-mini	0.9116	0.5507	**0.69**	0.73	0.07	0.00	0.00	0.12
VLM	Qwen-2.5-vl-72b	0.6859	0.4213	0.25	0.57	0.09	0.00	0.09	0.00
VLM	Gemma-3-27b	0.6929	0.3550	0.24	0.48	0.15	0.00	0.05	0.00
VLM + few shot	GPT-4.1	**0.9267**	0.4261	0.47	0.52	0.00	0.00	0.30	0.04
VLM + few shot	GPT-4.1-mini	0.8904	0.4940	0.46	0.62	0.12	0.00	0.30	0.00
VLM + few shot	Qwen-2.5-vl-72b	0.7464	0.3630	0.35	0.46	0.10	0.13	0.14	0.03
VLM + few shot	Gemma-3-27b	0.8083	0.3827	0.05	0.51	0.22	0.00	0.17	0.00
cVL	CLIP	0.6839	0.0169	0.19	0.00	0.00	0.00	0.00	0.04
cVL	BiomedCLIP	0.8868	0.2774	0.56	0.29	0.17	0.00	0.21	0.04

Table 6. Comparative analysis of machine learning models in polyp detection and Classification. Performance comparison of classical machine learning (CML) models, ResNet-50, vision Language models (VLMs), and specialized VLMs for polyp detection and classification tasks. The bolded values represent the highest F1 scores for each task in the column.

The most challenging classifications were observed for tubulovillous adenoma (TVA, 48 training, 11 test), villous adenoma (VA, 30 training, 6 test), and inflammatory polyp (IP, 38 training, 7 test) images. For TVA, ResNet50 achieved the highest F1 of 0.55, with Decision Tree (F1: 0.27) and SVM (F1: 0.25) showing limited effectiveness. Most other models, including VLMs and contrastive VLMs, performed at or near random chance, except for BiomedCLIP (F1: 0.17), Gemma-3-27b (F1: 0.15), and GPT-4.1-mini (F1: 0.07), which provided small improvements. For VA, ResNet50 (F1: 0.25) was the only model with moderate performance; most other models failed, with minor gains from Claude-3-Opus (F1: 0.04), GPT-4.1 (F1: 0.20), and Qwen-2.5-vl-72b (F1: 0.09). For IP, ResNet50 (F1: 0.71) performed best, followed by SVM (F1: 0.36) and Logistic Regression (F1: 0.20), while most VLMs were ineffective, except GPT-4.1-mini (F1: 0.12) and GPT-4.1 (F1: 0.08); contrastive models CLIP and BiomedCLIP (F1: 0.04 each) contributed minimally.

Figure 4 displays confusion matrices for polyp classification utilizing Random Forest (CML's top performer), ResNet50, GPT-4.1 (the leading VLM), and BiomedCLIP. Adenoma subtypes showed substantial confusion across all models, with tubulovillous and villous adenomas frequently misclassified as tubular adenomas. ResNet50 demonstrated the best discrimination but still showed considerable uncertainty. Complete ROC curves and confusion matrices for all models are in **Supplementary Figures S3 and S4**.

Polyp classification performance (CADx) on external validation dataset

External validation on 75 images from the PICCOLO database revealed varying performance degradation across model types. ResNet50 showed the largest decline (internal: 0.83, external: 0.49, Δ = -0.34), suggesting overfitting to institution-specific characteristics. VLMs demonstrated smaller drops: GPT-4.1-mini (0.75 to 0.59, Δ = -0.16), GPT-4.1 (0.72 to 0.58, Δ = -0.14), and Gemma-3-27B (0.72 to 0.53, Δ = -0.19). Notably, Qwen-2.5-vl-72B exhibited the smallest decline among high-performing models (0.66 to 0.61, Δ = -0.05), suggesting superior cross-institutional generalization. CML models showed intermediate degradation: SVM (0.69 to 0.52, Δ = -0.17), Logistic Regression (0.59 to 0.48, Δ = -0.11), Random Forest (0.63 to 0.53, Δ = -0.10), and Decision Tree (0.55 to 0.53, Δ = -0.02). Gaussian Naive Bayes showed apparent improvement (0.08 to 0.12, Δ = +0.04), likely reflecting statistical noise given its poor baseline. These results suggest that while CNN achieves superior internal performance, pretrained VLMs may offer generalization advantages. F1 scores are presented in Table 7, with confusion matrices provided in **Supplementary Figure S5**.

Exploring Few-Shot injection impact on VLM prediction

Performance of Few-shot prompting produced heterogeneous effects for polyp detection (F1 scores in Table 6; confusion matrices in **Supplementary Figure S6**). Gemma-3-27B showed the largest improvement (F1: 0.69 to 0.81), followed by Qwen-2.5-VL-72B (F1: 0.69 to 0.75). GPT-4.1 exhibited only a marginal gain (F1: 0.92 to 0.93), suggesting near-optimal baseline performance, while GPT-4.1-mini experienced a slight decline (F1: 0.91 to 0.89).

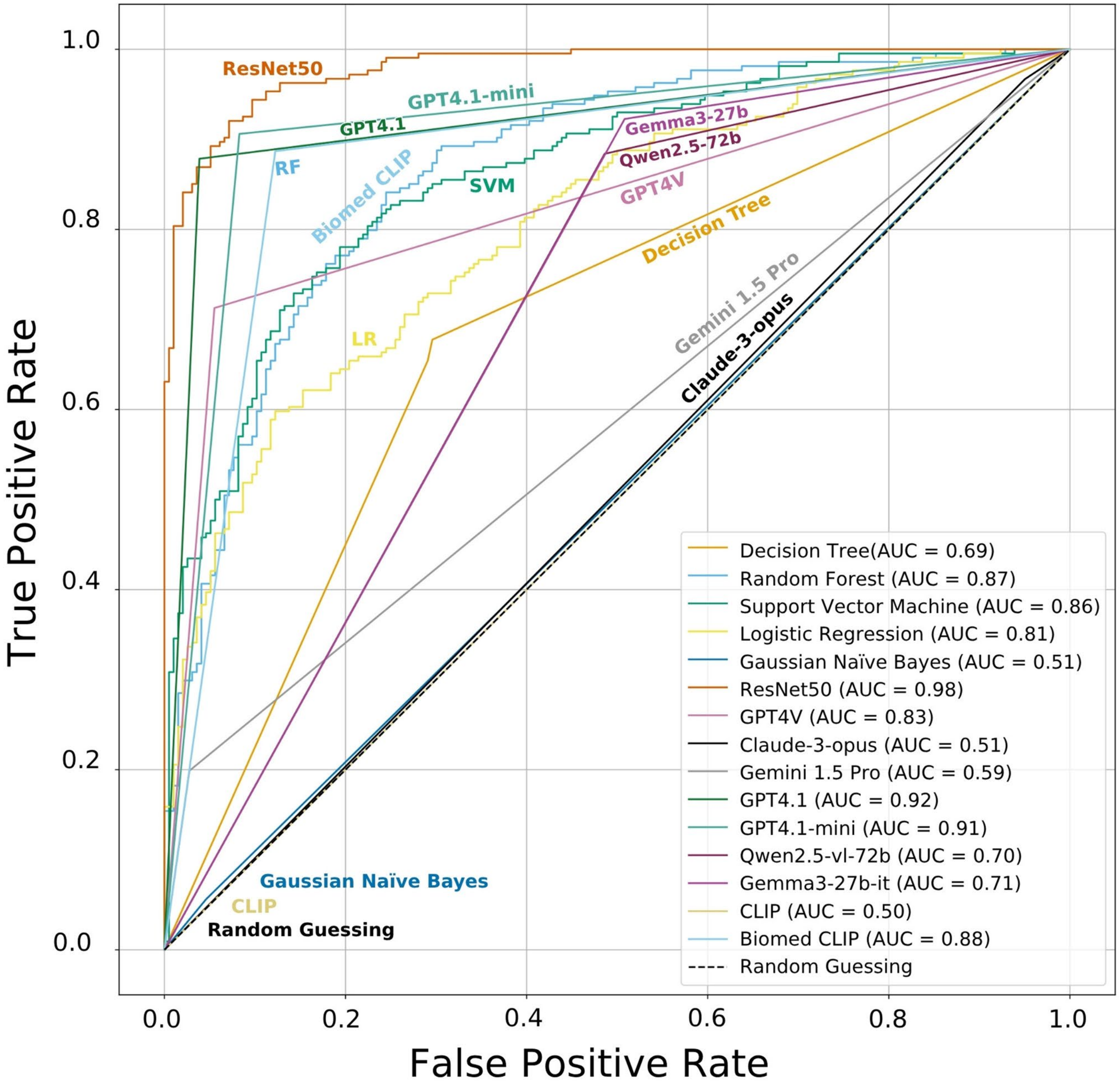

Fig. 2. ROC curves and AUROC values for polyp detection. Receiver operating characteristic curves for polyp detection, with the corresponding AUROC values. AUROC values greater than 0.8 are shown in bold.

Few-shot prompting also produced mixed effects on classification performance across models. While overall weighted F1 often declined (GPT-4.1: 0.55 to 0.43, GPT-4.1-mini: 0.55 to 0.49, Qwen-2.5-vl-72b: 0.42 to 0.36), certain underrepresented categories benefited substantially. For example, GPT-4.1-mini improved HP classification F1 score from 0.00 to 0.30, and Qwen-2.5-vl-72b increased AC from 0.25 to 0.35 and VA from 0.00 to 0.13. Gemma-3-27b also demonstrated consistent gains, raising weighted F1 from 0.36 to 0.38, with HP classification F1 score improving from 0.05 to 0.17 and TVA from 0.15 to 0.22. However, these improvements were often offset by declines in high-prevalence classes such as AC and TA (e.g., GPT-4.1 F1 score for AC: 0.61 to 0.47, TA: 0.71 to 0.52). This trade-off suggests few-shot learning requires careful calibration, as improvements for rare classes may come at the cost of common category accuracy.

Discussion

Our systematic evaluation established a performance hierarchy across computational paradigms. For polyp detection, the highest-performing zero-shot VLMs achieved parity with task-specific CNN. GPT-4.1 (F1: 91.98%) and GPT-4.1-mini (91.16%) performed comparably to ResNet50 (91.35%), demonstrating that frontier VLM architectures can match specialized CNN for binary classification tasks. The 11-percentage-point improvement from GPT-4 (81.02%) to GPT-4.1 within a single model generation suggests rapid architectural evolution,

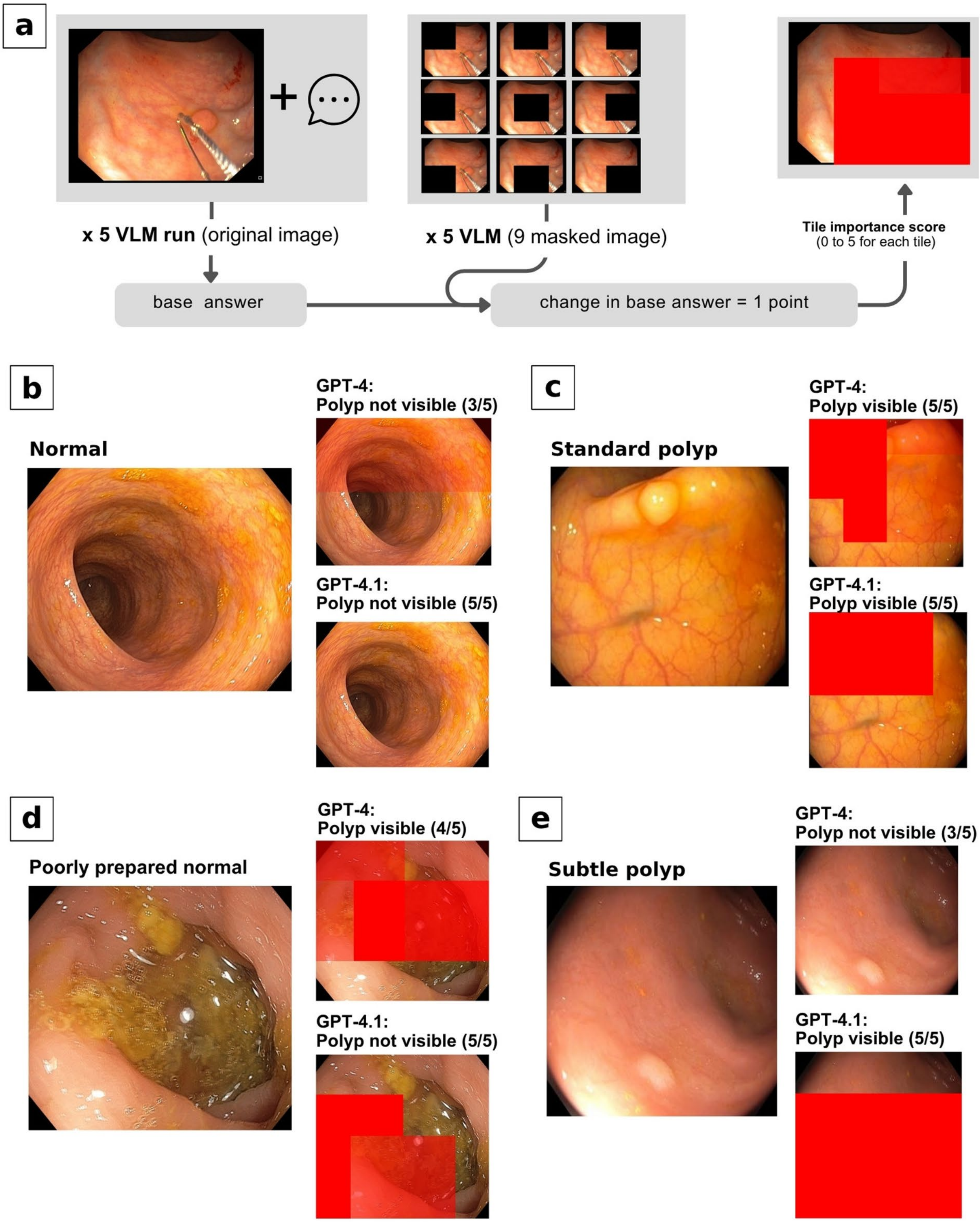

Fig. 3. Tile-level importance analysis of GPT-4.1 and GPT-4 polyp detection using TiLense. Evaluation of GPT-4.1 and GPT-4 for polyp detection using TiLense, focusing on tile-level importance. The method includes five runs with vision-language models (VLMs) on original and masked images, using 9 masked tiles per image. Each tile receives an importance score from 0 to 5, indicated by a color gradient from white to red, where red denotes a tile whose removal alters the base answer significantly. A reference answer for each image is established, and deviations are scored as 1 point. The final answer was considered by voting among five answers. Panels (**a**–**e**) show tile-level predictions across image conditions: standard image without polyp (**b**), standard image with polyp (**c**), challenging image without polyp and poor preparation (**d**), and challenging image with hard-to-see polyp (**e**).

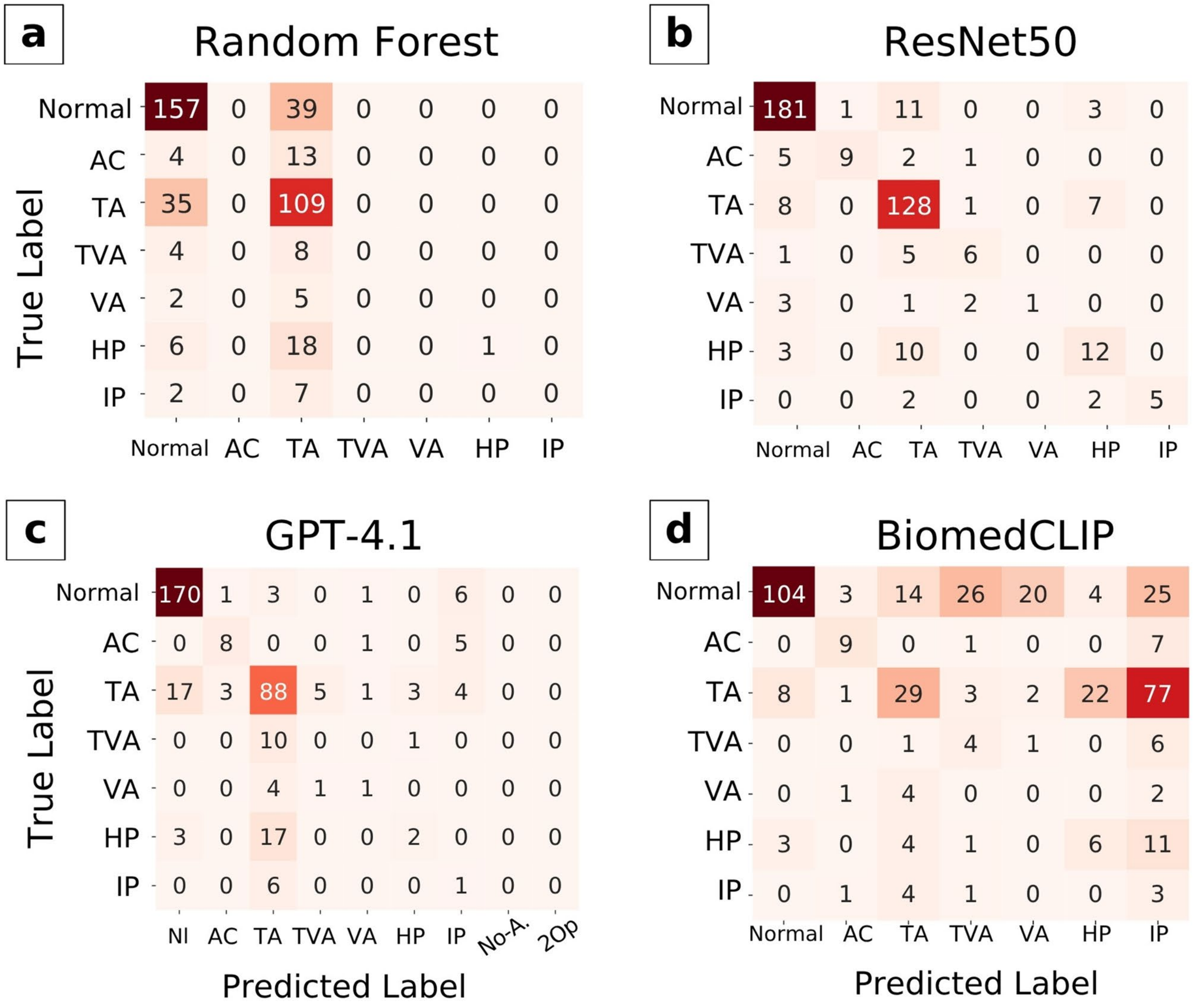

Fig. 4. Polyp classification performance of top-performing models. Confusion matrices of polyp classification are provided for the top-performing classical machine learning model (**a**: Random Forest), convolutional neural network (**b**: ResNet-50), highest-performing vision-language model (**c**: GPT-4.1), and the contrastive vision-language encoder fine-tuned on external general medical imaging data (**d**: BiomedCLIP). Abbreviations: AC, Adenocarcinoma; TA, Tubular Adenoma; TVA, Tubulovillous Adenoma; VA, Villous Adenoma; HP, Hyperplastic Polyp; IP, Inflammatory Polyp; No-A: No answer provided; 2OP: two options (polyp type) were selected.

though proprietary models preclude definitive attribution. However, this performance was not universal across VLMs. Qwen-2.5-vl-72b (68.59%), Gemma-3-27b (69.29%), Claude-3-Opus (66.40%), and Gemini-1.5-Pro (19.37%) performed substantially worse, with some scoring at or below CMLs baselines (Random Forest: 81.02%, SVM: 77.92%). This 72-point performance range across VLMs (GPT-4.1: 91.98% to Gemini-1.5-Pro: 19.37%) underscores that VLM does not denote uniform capability, but rather encompasses architectures with markedly different medical imaging performance.

For polyp classification, even the highest-performing VLMs underperformed CNN. ResNet50 (weighted F1: 74.94%) substantially outperformed GPT-4.1-mini (55.07%), the best VLM for this task. This 20-point performance gap widened substantially for rare polyp subtypes, as detailed below. CML approaches consistently underperformed deep learning methods for both detection and classification, validating the shift toward neural architectures in medical imaging.

This detection-classification dichotomy likely reflects fundamental task differences. Polyp detection requires distinguishing abnormal mucosal protrusions from normal tissue based on features such as texture variations, color changes, and surface irregularities visible during endoscopy. VLMs' broad pretraining on diverse visual domains may enable recognition of these general visual patterns. In contrast, polyp classification requires discrimination between subtle morphological variants visible on the polyp surface. Distinguishing different polyp classes based on colonoscopy images probably requires recognition of surface pit patterns, vascular patterns, color variations, shape characteristics, and surface texture that correlate with underlying histology[28,29]. These

Model Family	Model	Polyp Classification	AC ($N=79$)	A ($N=50$)	HP ($N=138$)	Polyp Classification	AC ($N=9$)	A ($N=50$)	HP ($N=16$)
		Test (Internal Validation)				External Validation			
		Weighted F1	F1	F1	F1	Weighted F1	F1	F1	F1
CML	Decision tree	0.55	0.16	0.64	0.27	0.53	0.00	0.73	0.22
CML	Random forest	0.63	0.00	0.78	0.07	0.53	0.00	**0.80**	0.00
CML	Support vector machine	0.69	0.38	0.78	0.31	0.52	0.00	0.78	0.00
CML	Logistic regression	0.59	0.11	0.72	0.07	0.48	0.00	0.65	0.25
CML	Gaussian naive bayes	0.08	0.09	0.08	0.07	0.12	0.00	0.07	**0.36**
CNN	ResNet50	**0.83**	0.71	**0.89**	**0.52**	0.49	0.62	0.53	0.32
VLM	GPT-4.1	0.72	0.64	0.83	0.14	0.58	0.59	0.77	0.00
VLM	GPT-4.1-mini	0.75	**0.72**	0.87	0.00	0.59	**0.75**	0.75	0.00
VLM	Qwen-2.5-vl-72b	0.66	0.37	0.76	0.16	**0.61**	0.55	0.71	0.32
VLM	Gemma-3-27b	0.72	0.50	0.84	0.12	0.53	0.30	0.66	0.27

Table 7. Comparative analysis of machine learning models in polyp classification in external Dataset. Performance comparison of classical machine learning (CML) models, ResNet-50 and vision Language models (VLMs) for polyp classification tasks. The bolded values represent the highest F1 scores for each task in the column. Abbreviations: CML, Classical Machine Learning; VLM, Vision Language Model; cVL, contrastive Vision-Language encoders; AC, Adenocarcinoma; A, Adenomatous; HP, Hyperplastic Polyp.

domain-specific visual-histological correlations, likely absent from general pretraining datasets, may explain why VLMs struggle with fine-grained histological prediction despite achieving strong detection performance.

Performance on rare polyp types revealed the magnitude of this classification limitation. For TA (650 training images, 121 test images), GPT-4.1 and GPT-4.1-mini achieved 71–73% F1 for endoscopic histological prediction. However, performance declined substantially for rarer subtypes: VA (30 training, 6 test) both models ≤ 20% F1; TVA (48 training, 11 test) both ≤ 7% F1; IP (38 training, 7 test) both ≤ 12% F1. For HP (116 training, 22 test), both achieved 0% F1. In contrast, ResNet50 maintained non-zero performance across all categories: HP 49%, VA 25%, TVA 55%, IP 71%. Even CML models (SVM: 31% for HP) outperformed the leading VLMs on these categories. This pattern extends beyond simple class imbalance, as classical models trained on the same limited rare examples maintained non-zero performance. The findings suggest that zero-shot transfer, while effective for common polyp types with abundant visual similarity to general pretraining data, fails for rare histological presentations requiring domain-specific pattern recognition.

The substantial performance variability across VLMs noted above warrants investigation. These findings are consistent with emerging evidence from other clinical domains showing wide variability in VLM performance across medical imaging tasks[30–35]. Several factors likely contribute. First, architectural differences across proprietary models affect visual-language integration. GPT-4.1-mini achieving nearly identical detection performance (91.16%) to GPT-4.1 (91.98%) despite presumably fewer parameters suggests architectural innovations rather than scale drive improvements. Second, pretraining data composition varies. BiomedCLIP (88.68% F1) substantially outperformed general CLIP (68.39%) for polyp detection as a result of its additional training on 15 million biomedical figure-caption pairs from PubMed Central[11], providing direct evidence that medical content exposure improves performance. General-purpose VLMs likely contain varying amounts of incidental medical imaging in their pretraining corpora, partially explaining performance differences. Third, instruction-following capability varies substantially, as demonstrated by our prompt engineering experiments.

Prompt engineering revealed substantial performance sensitivity. For polyp detection, improvements with engineered prompts ranged from 2.2% (GPT-4.1, Gemini-1.5-Pro) to 51.0% (Qwen-2.5-vl-72b). For classification, improvements were substantial: GPT-4.1 (15.6% to 59.4%, + 280.7%), GPT-4.1-mini (16.9% to 71.1%, + 320.7%), and Qwen-2.5-vl-72b (0.8% to 50.2%, + 6175%). These magnitudes underscore that systematic prompt design is critical for medical VLM deployment[17,36]. Few-shot prompting showed variable effects. For detection, Gemma-3-27B improved substantially (+ 17.4%) while GPT-4.1 showed minimal gain (+ 1.1%), consistent with baseline performance near ceiling. GPT-4.1-mini declined slightly (-2.2%), suggesting few-shot examples may introduce noise for high-performing models. This outcome may also be attributed to our selection of examples: we primarily included clear and unambiguous cases that the model could process effectively, whereas its performance may decline when confronted with more ambiguous images. For classification, few-shot prompting often improved rare categories while reducing common category performance, yielding limited overall gains. Our results exceed previously reported prompt-dependent performance variations and reinforce that effective prompt engineering is critical for clinical VLM implementation[17,36]. In addition, these findings reaffirm that prompt optimization benefits mid-performing models most, while top performers show diminishing returns[37–39].

Beyond internal performance patterns observed in our test set, cross-institutional generalization represents a critical consideration for clinical deployment. External validation on 75 images from the PICCOLO database assessed cross-institutional generalization. ResNet50 showed substantial performance decline (weighted F1: 0.83 to 0.49), potentially reflecting overfitting to institution-specific characteristics such as imaging equipment settings, acquisition protocols, or patient population differences. VLMs also experienced decreases, with GPT-4.1 (0.72 to 0.58), GPT-4.1-mini (0.75 to 0.59), and Gemma-3-27B (0.72 to 0.53) showing larger declines than Qwen-2.5-vl-72B (0.66 to 0.61). The relatively stable performance of some VLMs compared to ResNet50's larger degradation may suggest that zero-shot models pretrained on diverse data possess some cross-domain robustness. However, our limited external sample (75 images, one institution, three polyp classes versus six in internal data) precludes definitive conclusions.

These performance characteristics, together with fundamental differences in computational requirements, have direct implications for clinical deployment strategies. Computational requirements differ fundamentally between model families with direct implications for clinical applicability. CNNs require dataset annotation, model training (several hours on our hardware for ResNet50), and validation testing. However, once deployed, CNNs enable rapid local inference (milliseconds per image on CPU) with zero recurring costs and no network dependencies. This computational profile makes CNNs suitable for real-time intra-procedural applications, where frame-by-frame analysis during endoscope advancement can provide immediate feedback to endoscopists. VLMs eliminate training requirements through zero-shot deployment, substantially reducing barriers to entry. However, current API-based VLMs introduce per-image costs and network latency (seconds per image in our implementation), making them unsuitable for real-time use during live procedures. Network dependencies also introduce reliability concerns. The computational profile of current API-based VLMs restricts them to retrospective applications such as post-procedure quality assurance, batch analysis of stored images, or second-opinion consultation on challenging cases.

These computational constraints shape institutional deployment decisions. Academic centers with AI infrastructure may favor CNN development for real-time applications despite upfront costs, benefiting from zero marginal inference costs and real-time deployment capability for both detection and optical diagnosis. Community practices lacking machine learning expertise might find API-based VLMs useful for retrospective quality assurance despite recurring costs, as zero-training deployment enables immediate adoption for post-procedure review. However, institutions seeking real-time procedural guidance must pursue CNN-based approaches given current technological constraints. The substantial performance gap for rare polyp classification further indicates that current-generation VLMs should not be relied upon for optical diagnosis decisions without further technological advancement.

Several immediate research directions emerge from these findings. First, evaluation on video colonoscopy sequences rather than still frames would assess temporal reasoning capabilities and enable analysis of dynamic polyp characteristics across multiple viewing angles. Second, expansion of external validation to additional institutions with diverse endoscopy equipment, patient populations, and polyps would better characterize cross-institutional generalization and identify specific factors affecting model transferability. Third, investigation of spatial localization capabilities, particularly for VLMs through region-specific prompting or coordinate generation, would address a critical requirement for clinical applicability. Fourth, our choice of examples for few-shot prompting may have influenced the results; therefore, future studies should explore alternative methods for example selection. Finally, systematic analysis of model performance stratified by polyp size, morphology, and location would reveal potential biases affecting clinical safety and identify subgroups requiring targeted algorithmic improvements.

Several methodological limitations should be considered. First, natural prevalence disparities influenced our dataset composition despite our augmentation efforts, potentially impacting model performance for several rare polyp categories. Second, our evaluation used still colonoscopy images rather than video sequences, eliminating temporal continuity, polyp motion tracking, and multi-angle visualization available during actual procedures. Third, our study focuses on polyp detection (presence/absence) and classification (histological type) rather than spatial localization, which would be necessary for complete clinical implementation. Fourth, our external validation provides initial cross-institutional evidence but represents a small sample from a single additional institution with three polyp classes compared to our internal dataset's six classes. Larger-scale multi-institutional validation is necessary to establish robust generalizability benchmarks.

Conclusion

This systematic comparison of VLM and CNN for colonoscopy polyp analysis reveals a clear task-dependent performance hierarchy. While the highest-performing VLMs matched CNNs for binary polyp detection, CNNs maintained substantial advantages for polyp classification, particularly for rare polyp subtypes where VLMs failed entirely. These findings suggest that current zero-shot VLMs may serve retrospective quality assurance roles but remain unsuitable for real-time clinical deployment requiring histological discrimination. Computational constraints further restrict API-based VLMs to post-procedure applications, while CNNs enable real-time intra-procedural guidance. As both architectural families continue to evolve, understanding their complementary strengths and limitations will inform appropriate deployment strategies across diverse clinical settings.

Data availability

The datasets created and analyzed in this study cannot be accessed publicly due to IRB requirements; however, anonymized data can be obtained from the corresponding author (HAA) and SAASN ([sdamirsa@gmail.com]) upon request by providing the IRB code. The external dataset is accessible after signing data transfer agreement from [https://www.biobancovasco.bioef.eus/].The code for the generation and evaluation of responses is publicly available at: [https://github.com/aminkhalafi/CML-vs-LLM-on-Polyp-Detection].

Received: 28 March 2025; Accepted: 18 November 2025
Published online: 27 November 2025

References

1. Leufkens, A. M., van Oijen, M. G. H., Vleggaar, F. P. & Siersema, P. D. Factors influencing the miss rate of polyps in a back-to-back colonoscopy study. *Endoscopy* **44**, 470–475 (2012).
2. Kim, N. H. et al. Miss rate of colorectal neoplastic polyps and risk factors for missed polyps in consecutive colonoscopies. *Intest Res.* **15**, 411–418 (2017).
3. Rizkala, T., Menini, M., Massimi, D. & Repici, A. Role of artificial intelligence for colon polyp detection and diagnosis and colon cancer. *Gastrointest. Endosc. Clin. N. Am.* **35**, 389–400 (2025).
4. Pacal, I. et al. An efficient real-time colonic polyp detection with YOLO algorithms trained by using negative samples and large datasets. *Comput. Biol. Med.* **141**, 105031 (2022).
5. Karaman, A. et al. Robust real-time polyp detection system design based on YOLO algorithms by optimizing activation functions and hyper-parameters with artificial bee colony (ABC). *Expert Syst. Appl.* **221**, 119741 (2023).
6. Karaman, A. et al. Hyper-parameter optimization of deep learning architectures using artificial bee colony (ABC) algorithm for high performance real-time automatic colorectal cancer (CRC) polyp detection. *Appl. Intell.* **53**, 15603–15620 (2023).
7. Pacal, I. & Karaboga, D. A robust real-time deep learning based automatic polyp detection system. *Comput. Biol. Med.* **134**, 104519 (2021).
8. Ince, S., Kunduracioglu, I., Algarni, A., Bayram, B. & Pacal, I. Deep learning for cerebral vascular occlusion segmentation: A novel ConvNeXtV2 and GRN-integrated U-Net framework for diffusion-weighted imaging. *Neuroscience* **574**, 42–53 (2025).
9. Narasimha Raju, A. S. et al. CADxPolydetect: a clinically explainable hybrid deep learning system for multi-class colorectal lesion detection using augmented colonoscopy images. *BMC Med. Inf. Decis. Mak.* **25**, 335 (2025).
10. Radford, A. et al. Learning Transferable Visual Models From Natural Language Supervision. (2021).
11. Zhang, S. et al. BiomedCLIP: A multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. Preprint at (2025). https://doi.org/10.48550/arXiv.2303.00915
12. OpenAI, Achiam, J. & Adler, S. & others. GPT-4 Technical Report. (2024).
13. The Claude 3 Model Family. Opus, Sonnet, Haiku. in (2024).
14. Team, G. et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. (2024).
15. Pillai, A., Parappally, B. S. & Hardin, M. J. Evaluating the Diagnostic and Treatment Recommendation Capabilities of GPT-4 Vision in Dermatology. in *medRxiv* (2024). https://doi.org/10.1101/2024.01.24.24301743
16. Laohawetwanit, T., Namboonlue, C. & Apornvirat, S. Accuracy of GPT-4 in histopathological image detection and classification of colorectal adenomas. *J. Clin. Pathol.* **78**, 202–207 (2025).
17. Chen, R. et al. GPT-4 Vision on Medical Image Classification -- A Case Study on COVID-19 Dataset. Preprint at (2023). https://doi.org/10.48550/ARXIV.2310.18498
18. Han, T. et al. Comparative Analysis of GPT-4Vision, GPT-4 and Open Source LLMs in Clinical Diagnostic Accuracy: A Benchmark Against Human Expertise. Preprint at (2023). https://doi.org/10.1101/2023.11.03.23297957
19. Xu, P., Chen, X., Zhao, Z. & Shi, D. Unveiling the clinical incapabilities: a benchmarking study of GPT-4V(ision) for ophthalmic multimodal image analysis. *Br. J. Ophthalmol.* **108**, 1384–1389 (2024).
20. Yang, Z. et al. Performance of Multimodal GPT-4V on USMLE with Image: Potential for Imaging Diagnostic Support with Explanations. Preprint at (2023). https://doi.org/10.1101/2023.10.26.23297629
21. Jin, Q. et al. Hidden flaws behind expert-level accuracy of multimodal GPT-4 vision in medicine. *Npj Digit. Med.* **7**, 190 (2024).
22. Klement, W. & Emam, K. E. Consolidated reporting guidelines for prognostic and diagnostic machine learning modeling studies: development and validation. *J. Med. Internet. Res.* **25**, e48763 (2023).
23. Collins, G. S. et al. TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* **385**, e078378 (2024).
24. Haumaier, F., Sterlacci, W. & Vieth, M. Histological and molecular classification of Gastrointestinal polyps. *Best Pract. Res. Clin. Gastroenterol.* **31**, 369–379 (2017).
25. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 770–778 (2016). 770–778 (2016). (2016). https://doi.org/10.1109/CVPR.2016.90
26. OpenAI. GPT-4V(ision) System Card. in. (2023).
27. Safavi-Naini, S. A. A. et al. Vision-Language and Large Language Model Performance in Gastroenterology: GPT, Claude, Llama, Phi, Mistral, Gemma, and Quantized Models. Preprint at (2024). https://doi.org/10.48550/ARXIV.2409.00084
28. Sánchez-Montes, C. et al. Computer-aided prediction of polyp histology on white light colonoscopy using surface pattern analysis. *Endoscopy* **51**, 261–265 (2019).
29. Li, M. et al. Kudo's pit pattern classification for colorectal neoplasms: a meta-analysis. *World J. Gastroenterol.* **20**, 12649–12656 (2014).
30. Schmidl, B. et al. Assessing the use of the novel tool Claude 3 in comparison to ChatGPT 4.0 as an artificial intelligence tool in the diagnosis and therapy of primary head and neck cancer cases. *Eur. Arch. Otorhinolaryngol.* **281**, 6099–6109 (2024).
31. Nguyen, C., Carrion, D. & Badawy, M. K. Comparative performance of anthropic Claude and openai GPT models in basic radiological imaging tasks. *J. Med. Imaging Radiat. Oncol.* **69**, 431–439 (2025).
32. Ishida, M. et al. Diagnostic performance of GPT-4o and Claude 3 opus in determining causes of death from medical histories and postmortem CT findings. *Cureus* **16**, e67306 (2024).
33. Liu, X. et al. Claude 3 opus and ChatGPT with GPT-4 in dermoscopic image analysis for melanoma diagnosis: comparative performance analysis. *JMIR Med. Inf.* **12**, e59273 (2024).
34. Liu, M. et al. Evaluating the effectiveness of advanced large Language models in medical knowledge: A comparative study using Japanese National medical examination. *Int. J. Med. Inf.* **193**, 105673 (2025).
35. Chen, Z. et al. Assessing the feasibility of ChatGPT-4o and Claude 3-Opus in thyroid nodule classification based on ultrasound images. *Endocrine* **87**, 1041–1049 (2025).
36. Patil, R., Heston, T. F. & Bhuse, V. Prompt engineering in healthcare. *Electronics* **13**, 2961 (2024).
37. Chatterjee, A., Renduchintala, H. S. V. N. S. K., Bhatia, S. & Chakraborty, T. P. O. S. I. X. A Prompt Sensitivity Index For Large Language Models. Preprint at (2024). https://doi.org/10.48550/ARXIV.2410.02185
38. Zhuo, J. et al. ProSA: Assessing and Understanding the Prompt Sensitivity of LLMs. Preprint at (2024). https://doi.org/10.48550/ARXIV.2410.12405
39. Razavi, A. et al. Benchmarking Prompt Sensitivity in Large Language Models. Preprint at (2025). https://doi.org/10.48550/ARXIV.2502.06065

Acknowledgements

The authors utilized ChatGPT, Claude, and Grammarly to assist with manuscript editing. All authors have reviewed the manuscript and take full responsibility for its content.We want to particularly acknowledge the patients and the Basque Biobank integrated in the Platform ISCIII Biomodels and Biobanks (PT23/00013) for their collaboration.

Author contributions

MAK: Conceptualization, Methodology, Software, Investigation, Data Curation, Writing - Original Draft, Project administration, Visualization. SAASN: Methodology, Software, Writing - Original Draft, Project administration. AmSa: Investigation, Writing - Review & Editing. NN: Software, Writing - Review & Editing. DA: Validation, Writing - Original Draft, Visualization. PKM: Conceptualization, Data Curation. KK: Methodology, Supervision. NG: Methodology. SF: Validation, Writing - Review & Editing. SS: Investigation, Validation, Writing - Review & Editing. JSS: Investigation, Writing - Review & Editing. NPT: Validation, Writing - Review & Editing. NH: Writing - Review & Editing. GN: Validation, Resources, Writing - Review & Editing. HAA: Conceptualization, Methodology, Resources, Writing - Review & Editing, Supervision. AlSo: Conceptualization, Methodology, Validation, Resources, Writing - Original Draft.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-29566-2.

Correspondence and requests for materials should be addressed to H.A.A. or A.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.