

Multimodal Clinical Decision Support Using Vision-Language Models for Robust Gastrointestinal Disease Diagnosis and Medical Report Generation

Weicheng Shao

Department of Computer Science, University of Houston, Houston, TX, USA.
helloweicheng@uh.edu

Abhay Subramanian

Department of Electrical Engineering and Computer Science, University of Missouri,
Columbia, MO, USA.
helloabhay@missouri.edu

Bilal Bailey

Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence,
KS, USA.
baileybilal95@ku.edu

Abstract

Multimodal clinical decision support systems are emerging as a promising response to the fragmented information environment of gastrointestinal diagnosis. Endoscopic findings, patient history, laboratory data, and prior procedural notes must be interpreted together, yet most existing systems address only isolated tasks such as polyp detection or text classification. This paper examines the design of vision-language models for robust gastrointestinal disease diagnosis and medical report generation from a systems perspective. The analysis emphasizes structural trade-offs in multimodal integration, clinical reasoning, perceptual verification, language generation, and deployment infrastructure. A central argument is that diagnostic safety depends less on benchmark accuracy alone and more on the ability of the system to maintain calibrated behavior under input variability, documentation heterogeneity, and device-induced visual drift. The paper discusses architectural choices including modular encoders, retrieval-augmented clinical reasoning, structured report generation, and asynchronous deployment. It further addresses fairness, bias, governance, and sustainability as first-order design constraints rather than post hoc considerations. The proposed perspective connects visual segmentation, large language modeling, and institutional governance into a coherent framework for responsible clinical translation. Future directions include longitudinal multimodal learning, federated governance mechanisms, and prospective clinical evaluation that accounts for report quality, workflow integration, and patient safety.

Keywords

clinical decision support; vision-language models; gastrointestinal disease diagnosis; medical report generation; multimodal learning; robustness; algorithmic governance.

1. Introduction

The transformation of medical artificial intelligence from narrow image classifiers toward multimodal systems has created new possibilities for gastrointestinal diagnosis. Attention-

based architectures originally demonstrated the capacity to model long-range dependencies in language [1]. Subsequent bidirectional language models enabled robust clinical text representation [2], while visual transformers extended attention mechanisms to image understanding [3]. Contrastive language-image pretraining established shared representations between visual and textual modalities [4]. More recently, models that learn to combine frozen image encoders with large language models [5] and visual instruction tuning [6] have made it possible to pursue systems that interpret endoscopic images and generate clinically meaningful text. These technical developments coincide with a growing recognition that gastrointestinal decision making requires simultaneous interpretation of mucosal appearance, patient symptoms, prior interventions, and histopathological context.

The clinical problem addressed in this paper is the design of a multimodal clinical decision support system for robust gastrointestinal disease diagnosis and automatic medical report generation. A useful system must recognize polyps, inflammatory changes, bleeding, anatomical landmarks, and other findings while producing coherent reports that support endoscopists, referring providers, and quality assurance workflows. Unimodal systems often fail when visual evidence is ambiguous, when textual descriptions are inconsistent, or when input variability arises from different endoscopic devices, bowel preparation quality, and documentation practices. Large language models have shown substantial clinical knowledge [7], but their use in medicine introduces new risks related to fluency without factual grounding [8]. Foundation models for generalist medical artificial intelligence have been proposed as a route toward more flexible clinical reasoning because they can integrate multiple data streams within a shared semantic space [9]. However, the translation of these capabilities into gastrointestinal practice requires careful attention to system architecture, data governance, robustness, and clinical accountability [10].

This paper adopts a systems-level orientation rather than presenting a single model implementation. It emphasizes structural trade-offs, deployment constraints, and institutional safeguards. The analysis draws on prior work in deep learning for healthcare [11], algorithmic fairness [12], and open gastrointestinal imaging datasets [13] to build a framework that connects perception, reasoning, and reporting. A central concern is that multimodal models may inherit the failure modes of both visual classifiers and generative language systems. A missed lesion can become an omitted finding, while a hallucinated phrase can become an inappropriate surveillance recommendation. Addressing these risks requires architectural separation between visual evidence and linguistic expression, as well as evaluation protocols that assess report safety rather than text similarity alone.

2. Background and Clinical Motivation

Gastrointestinal endoscopy generates high-dimensional visual data in which pathology is often indicated by subtle changes in mucosal architecture, vascular patterns, and lesion contours. Deep learning systems have achieved strong performance in selected diagnostic tasks, but many models are trained on narrow datasets and lack the semantic reasoning needed to connect visual findings with clinical narratives [11]. A clinical decision support tool for gastrointestinal disease cannot be reduced to a single polyp detector. It must address multiple disease categories, anatomical sites, severity grades, and surveillance implications while maintaining reliability across patient populations and acquisition settings. This requirement intersects with long-standing concerns about algorithmic bias in health care, where models trained on heterogeneous data can learn spurious associations that reproduce disparities [12]. For gastrointestinal imaging, open repositories such as HyperKvasir provide multi-class

endoscopic image and video data covering anatomical landmarks and pathological findings, but they also expose the challenge of combining large-scale visual data with structured clinical context [13].

The clinical motivation for multimodal reasoning is especially strong in gastroenterology. The same visual pattern may have different implications depending on bleeding risk, anticoagulant status, symptom duration, or prior endoscopic findings. A small polyp may be clinically insignificant in one patient and justify polypectomy in another. A suspicious lesion may require immediate biopsy when cancer is suspected, but a similar appearance in a patient with inflammatory bowel disease may represent regenerative tissue. These distinctions cannot be made by visual inspection alone. They require the integration of clinical history, medication use, family history, and laboratory findings. A vision-language system that jointly models imaging and text can, in principle, express these dependencies more explicitly than a purely visual classifier. However, this capability also introduces challenges related to input variability, report reliability, and the need for institutional governance.

The broader landscape of medical artificial intelligence has shifted from task-specific deep learning models toward foundation models that can be adapted across many clinical applications [9]. Despite this shift, most deployed systems remain narrow and operate within carefully bounded contexts. The difference between research demonstrations and clinical infrastructure is substantial. A model that performs well on a curated test set may fail when deployed in a different hospital because of changes in documentation style, patient demographics, or imaging equipment. The present work therefore treats robustness as a central design objective rather than an afterthought. It considers how architectural choices, data curation, prompt normalization, and monitoring mechanisms can make multimodal gastrointestinal decision support safer and more sustainable in real clinical environments.

3. System Architecture for Multimodal Data Integration

A vision-language clinical decision support system can be conceptualized as a distributed pipeline linking endoscopic image acquisition, de-identified data curation, feature encoding, cross-modal fusion, clinical reasoning, and report generation. The architecture should separate low-level perception from high-level semantic inference to allow independent update cycles. The visual encoder may be a transformer-based architecture pretrained on natural images and then adapted to endoscopic frames, while the language encoder is derived from a clinically oriented transformer [2], [3]. Contrastive visual-language alignment [4] and bootstrapped learnable querying [5] are attractive because they reduce the need for expensive dense annotations. Visual instruction tuning [6] provides a mechanism for aligning multimodal representations with clinical instructions such as lesion description, differential diagnosis, and recommendation generation. However, a production system must be more than a single model. It requires orchestration, versioning, audit logs, human review queues, and a clear separation between decision support and automated diagnosis.

A crucial trade-off in the system architecture concerns where clinical knowledge is stored. End-to-end vision-language models can produce fluent responses, but they may hallucinate disease findings or overlook rare conditions when the training distribution lacks sufficient gastrointestinal examples. An alternative governance-oriented architecture combines a retrieval module over clinical guidelines and prior reports with a generative model that is constrained by retrieved evidence. This design reduces the risk of unsupported assertions while preserving the flexibility of language generation. The choice has structural implications for latency, memory requirements, and model updating. In a hospital deployment, the system

may need to run on local infrastructure to avoid transmitting protected health information to external services. The architecture must therefore support modular deployment, with an encoder that can be frozen and updated separately, a reasoning module that can be audited, and a report generation layer that can be evaluated against institutional templates. Such modularity also enables sustainability because smaller domain-specific adapters can be retrained without repeating the full pretraining process.

The data integration layer deserves particular attention. Endoscopic images are rarely sufficient by themselves. The system must incorporate free-text indications, prior report summaries, medication lists, and relevant laboratory values without overwhelming the model with irrelevant detail. A common approach is to represent each data modality in a shared latent space and then apply cross-attention to allow the language model to select relevant visual and textual evidence [5], [6]. However, this approach can obscure the provenance of a generated statement. A safer design is to retain the source of each clinical fact and to require the model to reference that source when generating a report. This is not merely a technical preference. It supports auditability by allowing a clinician to trace a suspected finding back to an endoscopic frame or a specific phrase in the medical record. The system can also flag cases in which the visual and textual streams disagree, routing those cases for human review rather than forcing the model to resolve the conflict silently.

4. Vision-Language Clinical Reasoning and Diagnostic Trade-offs

Clinical reasoning in gastrointestinal disease differs from standard image classification because the same visual pattern may have different implications depending on patient context. A vision-language model can be prompted to produce a structured differential diagnosis rather than a single label. For example, a sessile polyp with surface irregularity may be described in visual terms, compared with guideline criteria, and linked to follow-up intervals. A purely visual classifier cannot easily articulate uncertainty about whether a lesion is hyperplastic or adenomatous. In this sense, language generation is not merely a convenience. It is a mechanism for making clinical reasoning explicit and therefore subject to review. Cross-domain comparisons with radiology are instructive, but gastrointestinal endoscopy introduces additional complexity because the visual field is dynamic, lighting is inconsistent, and important findings may appear only briefly in a video sequence.

The distinction between descriptive statement and diagnostic interpretation is essential. A well-designed system will separate the visual finding from the inferred pathology to avoid premature closure. It might state that a flat lesion with central depression was observed in the ascending colon and then provide a differential diagnosis rather than a definitive cancer label. This behavior can be reinforced through instruction tuning and through controlled decoding that keeps the model within predefined report sections. The structural trade-off is that excessive constraining may reduce fluency or omit uncommon but relevant observations, while insufficient constraining may produce confident language that is not supported by the visual record. The system should therefore be evaluated on calibrated confidence, refusal behavior, and the ability to flag incomplete examinations rather than on text similarity alone.

Another trade-off concerns the division of labor between the visual encoder and the language model. If the language model has direct access to raw visual embeddings, it may learn contextual associations that improve accuracy on common cases but degrade under distribution shift. If the visual encoder is instead required to produce a structured intermediate representation, such as lesion location, size, morphology, and likely pathology, then the language model can reason over a more stable symbolic input. This intermediate

representation can be verified independently and compared with the generated report. The cost is that some visually subtle findings may not fit neatly into the predefined categories. A hybrid design that combines structured visual attributes with a limited set of learned embeddings may balance interpretability and expressiveness. The key principle is that no model component should be allowed to generate a high-consequence diagnostic claim without an auditable evidence path.

5. Robustness Under Input Variability and Dataset Shift

A central limitation of many clinical language models is their sensitivity to input phrasing, document structure, and missing data. Recent work has systematically evaluated large language model performance under health care input variability and found measurable degradation when the same clinical information is expressed in different formats or when irrelevant detail is introduced [14]. This finding is directly relevant to gastrointestinal decision support, where endoscopy reports vary substantially across institutions, endoscopists, and documentation systems. A model that performs well on curated benchmarks may fail when confronted with abbreviation expansion, template variation, or changes in lesion nomenclature. Designing for robustness therefore requires data augmentation, prompt normalization, and consistency checks across multiple paraphrases of the same clinical query. It also requires monitoring model outputs across diverse deployment sites to detect drift in diagnostic sensitivity and report quality.

For the visual component, variability arises from endoscope manufacturer, illumination, preparation quality, and motion artifacts. Traditional detection models have been adapted to gastrointestinal polyp detection and segmentation to reduce false negatives during screening [15]. Segmentation datasets for polyps further enable the training of boundary-aware models that can highlight lesion extent and support follow-up decisions [16]. Yet the integration of these visual models with language reasoning introduces the possibility that a segmentation error may be amplified into a textual claim about pathology. A robust multimodal system must therefore pass visual evidence through multiple consistency filters before it is committed to a report. One approach is to require agreement between the segmentation mask, the visual encoder representation, and the clinical context before a high-confidence diagnostic phrase is generated. Another is to use calibrated uncertainty to express ambiguous visual findings through language such as suspected or cannot exclude. These design choices connect directly to patient safety because an incorrect polyp classification can change surveillance intervals and resource allocation.

Dataset shift is not only technical but also organizational. When a hospital changes endoscopy platforms or introduces new reporting templates, the statistical distribution of both images and text changes. A model that was validated on historical data may no longer be reliable. The system architecture should therefore include a monitoring layer that compares incoming data with the training distribution and alerts the clinical informatics team when drift exceeds a predefined threshold. This layer can track simple image statistics such as color distributions and lighting characteristics, as well as text features such as report length, abbreviation density, and symptom vocabulary. The goal is not to prevent all changes, but to make them visible so that model recalibration can be planned deliberately. Without such monitoring, silent degradation may go unnoticed until a serious diagnostic error occurs.

6. Medical Report Generation and Clinical Language Reliability

Automatic medical report generation requires a different evaluation lens than diagnostic accuracy. A report may be factually correct but clinically unhelpful if it omits anatomical location, size, morphology, or follow-up recommendation. In radiology, large annotated datasets have enabled the development of report generation systems that produce structured findings and impressions [17], [18]. In gastrointestinal imaging, the absence of comparably large paired endoscopic image-report datasets has slowed progress. Polyp appearance models developed for automated detection illustrate the importance of domain-specific visual descriptions, including shape, texture, and color cues that inform clinical assessment [19]. A vision-language report generator must learn to translate these visual properties into terminology that aligns with endoscopy reporting standards and downstream decision making.

The reliability of generated reports is a socio-technical challenge. A model may generate a grammatically fluent report that contains a subtle but clinically consequential error, such as describing a lesion in the cecum when it was actually in the sigmoid colon. Language models also tend to overstate diagnostic certainty when they are trained to maximize likelihood rather than to represent uncertainty. To mitigate these failures, the system can generate structured reports with mandatory fields for location, size, morphology, and follow-up. It can also compare generated statements against a separate visual classifier that predicts the same attributes independently. When disagreement occurs, the report can be routed for human review or softened to reflect uncertainty. This division of labor between perceptual verification and linguistic generation is essential for safe deployment. Report generation should be seen as an assistive task that summarizes evidence and proposes language, rather than as a replacement for the endoscopist's interpretive responsibility.

Clinical language reliability also depends on the reporting standards of the institution. Endoscopy reports often follow specialty-specific vocabularies that vary by region and by training background. A model that produces reports in a generic style may be less acceptable to clinicians even if the clinical content is correct. Involving gastroenterologists in the design of report templates and acceptable phrasing can improve adoption. The model can be tuned to follow institutional preferences while preserving factual consistency. This requires a careful balance between customization and generalization. Excessive local adaptation may reduce the transferability of the model, while excessive standardization may ignore meaningful clinical distinctions. The governance framework should therefore include a controlled process for template updates and language model versioning.

7. Fairness, Bias, and Algorithmic Governance

The deployment of vision-language clinical decision support raises governance questions that extend beyond technical accuracy. Clinical AI systems are embedded in institutional workflows, regulatory frameworks, and reimbursement structures, and their benefits are not evenly distributed unless access and performance are deliberately evaluated across populations [20]. Bias can enter through training data, label definitions, and the implicit assumptions encoded in clinical language. Because gastrointestinal disease prevalence and presentation vary by age, geography, and genetic background, a model trained on one population may systematically underperform on another. The governance challenge is to define acceptable performance across demographic and clinical subgroups and to monitor those metrics continuously. This requires an institutional data governance framework that documents training data provenance, update frequency, and exclusion criteria [21]. It also requires clear accountability when a model contributes to a missed lesion or an inappropriate surveillance recommendation.

Explainability is often proposed as a solution to bias, but current explanation methods may provide post hoc rationalizations that do not reliably reveal the causal factors behind a model decision [22]. In gastrointestinal diagnosis, an attention map over an endoscopic image may highlight the location of a lesion while obscuring the model's reliance on an irrelevant background feature. Therefore, governance should emphasize prospective testing, subgroup analysis, and human review rather than visual explanations alone. The system should maintain an audit trail of model versions, inputs, outputs, and clinician overrides. Regulatory submissions must include evidence of robustness under input variability, as well as performance on rare and high-consequence conditions. Without such governance, the adoption of multimodal models may accelerate clinical workflows while silently embedding new forms of diagnostic inequity.

8. Deployment Infrastructure and Sustainability

Deployment of a vision-language system in gastrointestinal practice requires balancing model expressiveness with computational cost. Large transformer models have high memory and energy requirements, which can conflict with the limited computational resources of outpatient endoscopy units. One structural solution is to use frozen pretrained encoders and lightweight adapters that are fine-tuned on institutional data [5], [6]. This approach reduces training energy and permits frequent updates without retraining the entire model. Another solution is to deploy the system as a microservice that processes images and drafts reports only when the endoscopist requests decision support. Such on-demand operation reduces idle energy consumption and avoids unnecessary data movement. Latency is nevertheless important because endoscopic procedures are time-sensitive, and a slow report generation step may disrupt workflow. The architecture should therefore support asynchronous preprocessing of video frames while the procedure is in progress, followed by rapid text generation at the time of report finalization.

Sustainability also includes the long-term maintenance of model performance. Clinical language models degrade when documentation practices change or when new endoscopic devices produce different color distributions. Continuous monitoring requires a feedback loop in which clinicians correct generated reports and flagged disagreements are stored in a de-identified improvement dataset. This feedback loop creates new governance obligations because corrections contain protected health information and because model updates can alter behavior in subtle ways. A change management board with clinical, technical, and ethics representation can review update requests and decide whether a model revision is warranted. This organizational structure prevents ad hoc modifications and ensures that model changes are aligned with patient safety and institutional policy.

9. Evaluation, Clinical Translation, and Future Directions

Clinical translation of multimodal decision support should follow a staged process that begins with retrospective validation on de-identified data, moves to shadow deployment in which model outputs are not shown to clinicians, and then proceeds to prospective evaluation with human review. The evaluation should include diagnostic metrics for lesion detection and segmentation [15], [16], but must also assess the quality and safety of generated reports. Metrics such as completeness, factual consistency, and clinical acceptability may be more informative than generic language similarity. Human evaluation panels should include gastroenterologists, pathologists, and endoscopy nurses because they observe different failure modes. A system that improves detection sensitivity but generates verbose or misleading reports may not be clinically desirable. Conversely, a system that writes concise reports but

misses subtle lesions may create false reassurance. The evaluation protocol must therefore combine visual task performance with language reliability and workflow integration.

Future directions include the development of longitudinal multimodal models that incorporate prior endoscopic findings, pathology results, and molecular markers to support surveillance decisions. There is also a need for federated learning frameworks that allow institutions to improve shared models without centralizing sensitive patient data. Such frameworks introduce additional security and incentive challenges, but they may reduce the representational gaps that arise when only high-resource centers contribute training data. The field should also invest in open multimodal benchmarks for gastrointestinal disease that include varied documentation styles and acquisition conditions, consistent with the broader call for rigorous evaluation under input variability [14]. Finally, regulatory science must evolve to address generative clinical text, because traditional approval pathways for fixed diagnostic devices do not fully capture the dynamic, language-producing behavior of vision-language models. A future certification regime may require continuous postmarket monitoring, versioned model reporting, and explicit limits on the model's role in diagnostic decision making.

10. Conclusion

This paper has presented a system-level analysis of multimodal clinical decision support using vision-language models for gastrointestinal disease diagnosis and report generation. The core argument is that technical accuracy alone is insufficient. Robust deployment requires careful architectural decomposition, domain-specific multimodal alignment, and governance structures that address input variability, bias, and clinical workflow integration. The system must balance the expressive power of generative language with the need for perceptual verification and auditability. Maintaining this balance requires that generated report contents remain demonstrably anchored to visual and textual inputs. The model should not be permitted to produce a high-consequence diagnostic phrase unless the corresponding visual segmenter or detector has produced an aligned finding, or unless the clinical text contains explicit evidence that was retrieved and shown to the model. This evidentiary constraint should be enforced at the system level rather than left to prompting alone. In practice, this means that report generation functions as a structured summary of verified observations, not as an open-ended continuation of a clinical prompt. Such a design reduces the frequency of unsupported lesion descriptions and makes disagreements between perception and language available for human review.

Furthermore, the regulatory and institutional conditions for deployment deserve as much attention as model architecture. A vision-language system that is allowed to operate without clear performance thresholds, without subgroup monitoring, and without a mechanism for clinician correction may produce a false sense of confidence. Institutional review boards, data protection officers, and clinical informatics teams must therefore be integrated into the lifecycle of the system. Model updates should be documented in a versioned registry, and any change in report style or diagnostic behavior should trigger a revalidation cycle. This perspective aligns with the need to treat clinical AI as a continuously evaluated intervention rather than as a fixed software release.

Future research should develop prospective multimodal benchmarks that measure report factual consistency, unnecessary alarm generation, and endoscopist workflow disruption. Studies should also examine whether assistive report generation reduces documentation time without reducing the completeness of the procedural note. Finally, the field should continue to

investigate federated and privacy-preserving training methods that enable smaller centers to contribute diverse gastrointestinal data without centralizing protected health information. Only through such a combined technical and organizational effort can multimodal decision support realize its potential for safer, more equitable gastrointestinal care.

References

1. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* 30 (pp. 5998–6008). Curran Associates.
2. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 4171–4186). Association for Computational Linguistics.
3. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.
4. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning* (pp. 8748–8763). PMLR.
5. Li, J., Li, D., Savarese, S., & Hoi, S. (2023). BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning* (pp. 19730–19742). PMLR.
6. Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023). Visual instruction tuning. *Advances in Neural Information Processing Systems*, 36, 34892–34916.
7. Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Babiker, A., Schärli, N., Chowdhery, A., Mansfield, P., Demner-Fushman, D., & Natarajan, V. (2023). Large language models encode clinical knowledge. *Nature*, 620(7972), 172–180.
8. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12), Article 248.
9. Moor, M., Banerjee, O., Abad, Z. S. H., Krumholz, H. M., Leskovec, J., Topol, E. J., & Rajpurkar, P. (2023). Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956), 259–265.
10. Rajpurkar, P., Chen, E., Banerjee, O., & Topol, E. J. (2022). AI in health and medicine. *Nature Medicine*, 28(1), 31–38.
11. Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., & Dean, J. (2019). A guide to deep learning in healthcare. *Nature Medicine*, 25(1), 24–29.

12. Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447–453.
13. Borgli, H., Thambawita, V., Smedsrud, P. H., Hicks, S., Jha, D., Eskeland, S. L., Randel, K. R., Pogorelov, K., Lux, M., Nguyen, D. T. D., Johansen, D., Griwodz, C., Stensland, H. K., Garcia-Ceja, E., Schmidt, P. T., Hammer, H. L., Riegler, M. A., Halvorsen, P., & de Lange, T. (2020). HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. *Scientific Data*, 7, Article 283.
14. Joshi, S., Mehta, M., Maniar, S., Wang, M., & Singh, V. K. (2026). Performance of large language models under input variability in health care applications: dataset development and experimental evaluation. *JMIR AI*, 5, e83640.
15. Tang, Y., Chang, C., Zhao, Z., Gu, Z., & Li, Z. (2026, March). Intelligent Detection and Segmentation of Gastrointestinal Polyps Based on YOLOv11 Deep Learning. In 2026 9th International Conference on Advanced Algorithms and Control Engineering (ICAACE) (pp. 825-828). IEEE.
16. Jha, D., Smedsrud, P. H., Riegler, M. A., Halvorsen, P., de Lange, T., Johansen, D., & Johansen, H. D. (2020). Kvasir-SEG: A segmented polyp dataset. In *MultiMedia Modeling: 26th International Conference, MMM 2020* (pp. 451–462). Springer.
17. Johnson, A. E. W., Pollard, T. J., Berkowitz, S. J., Greenbaum, N. R., Lungren, M. P., Deng, C.-Y., Mark, R. G., & Horng, S. (2019). MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6, Article 317.
18. Jing, B., Xie, P., & Xing, E. (2018). On the automatic generation of medical imaging reports. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (pp. 2577–2586). Association for Computational Linguistics.
19. Bernal, J., Sánchez, J., & Vilariño, F. (2012). Towards automatic polyp detection with a polyp appearance model. *Pattern Recognition*, 45(9), 3166–3182.
20. Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine*, 25(1), 44–56.
21. Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., Bakas, S., Galtier, M. N., Landman, B. A., Maier-Hein, K., Ourselin, S., Sheller, M., Summers, R. M., Trask, A., Xu, D., Baust, M., & Cardoso, M. J. (2020). The future of digital health with federated learning. *NPJ Digital Medicine*, 3, Article 119.
22. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215.