ESP Journal of Engineering Technology Advancements [ESPJETA]
International Conference on Emerging Computational Technologies
Eternal Scientific Publication / Conference Proceedings - (ICECT-2024) / ISSN: 2583-2646

Original Article

Explainable AI for Polyp Detection: Interpreting Deep Learning Models in Colorectal Cancer Diagnosis

Kamal Mundey
RC College of Engineering, Banagalore, India.

Abstract: *Colorectal cancer (CRC) remains one of the leading causes of cancer-related deaths globally, with early detection playing a pivotal role in improving patient prognosis. Polyps, which can develop into cancer, are often detected through colonoscopy or other medical imaging techniques. Recent advancements in deep learning have shown promising results in automating polyp detection, offering the potential to aid healthcare professionals in identifying at-risk patients. However, the lack of interpretability in deep learning models presents a significant challenge for their widespread adoption in clinical settings. This research focuses on developing an explainable AI framework for polyp detection, aimed at enhancing the transparency and trustworthiness of deep learning models in the context of colorectal cancer diagnosis. By applying state-of-the-art explainability techniques such as Grad-CAM, SHAP, and LIME, the study seeks to provide meaningful insights into how deep learning models arrive at their decisions, thus allowing clinicians to validate and understand model predictions. Through a comprehensive evaluation of model performance, explainability, and clinician feedback, this study aims to bridge the gap between cutting-edge AI technology and its practical, trustworthy application in the healthcare domain.*

Keywords: *Colorectal Cancer, Polyp Detection, Deep Learning, Explainable AI, Model Interpretability, Medical Imaging, Convolutional Neural Networks (CNN), Grad-CAM, LIME, SHAP, Trustworthy AI, Healthcare Applications, Colonoscopy, Artificial Intelligence In Medicine, Cancer Diagnosis, AI Transparency.*

I. INTRODUCTION

Colorectal cancer (CRC) is one of the most prevalent and deadly cancers worldwide, with millions of new diagnoses and deaths annually. The detection of colorectal polyps, which are precursors to cancer, plays a critical role in reducing the incidence of CRC. Colonoscopy, CT colonography, and other imaging techniques are routinely used for the identification of these polyps. However, manual examination of these images is time-consuming and subject to human error, which has prompted the integration of artificial intelligence (AI) to assist in polyp detection. Deep learning models, particularly Convolutional Neural Networks (CNNs), have emerged as powerful tools for automating polyp detection with high accuracy. Despite their success, these models are often seen as "black boxes," meaning their decision-making processes are not transparent or understandable to clinicians. This lack of interpretability is a significant barrier to the adoption of AI in clinical practice, as healthcare professionals must trust the system's outputs to make life-altering decisions. Explainable AI (XAI) offers a potential solution by providing a way to interpret and explain the decisions made by complex models. In the context of polyp detection for CRC diagnosis, the ability to explain AI predictions could enhance clinician confidence, leading to broader acceptance and improved outcomes. This research aims to develop and evaluate an explainable AI framework specifically designed for polyp detection in colorectal cancer diagnosis.

II. LITERATURE REVIEW

The detection of colorectal polyps has long been a critical task in the prevention and early diagnosis of CRC. Traditional methods for polyp detection, such as colonoscopy, have several limitations, including operator dependency, fatigue, and the potential for missed diagnoses. In recent years, deep learning models, particularly CNNs, have been increasingly applied to the task of polyp detection. These models excel at identifying subtle patterns in large, complex datasets like medical images. Several studies have demonstrated the potential of AI-based systems to outperform human experts in terms of accuracy, sensitivity, and speed. However, the complexity of these models means they often lack transparency, making it difficult for clinicians to understand how or why a certain diagnosis was made. This is especially problematic in healthcare, where understanding the reasoning behind a model's decision is essential for clinical trust and decision-making. To address this challenge, researchers have developed various techniques for explaining deep learning models, such as Grad-CAM, LIME (Local Interpretable Model-agnostic Explanations), and SHAP (Shapley Additive Explanations). These techniques aim to provide human-readable explanations of AI predictions by highlighting the features that influenced the model's output. Although these methods have been successfully applied in other domains, their use in medical imaging for polyp detection is still emerging. This section reviews the existing literature on polyp detection, deep learning in medical image analysis, and the techniques used to interpret AI models, identifying the gaps and challenges that this research aims to address.

III. PROBLEM STATEMENT

The advent of AI-based tools for medical image analysis has the potential to revolutionize the detection of colorectal polyps, leading to earlier diagnosis and improved patient outcomes. However, a significant challenge remains in ensuring that these tools are both accurate and interpretable. Deep learning models, particularly CNNs, have achieved impressive performance in polyp detection, but they often operate as "black-box" systems, offering little insight into the decision-making process. In a clinical context, this lack of transparency is problematic because it undermines clinician trust in the system. For a deep learning model to be adopted in routine clinical practice, healthcare professionals must understand and be able to validate its predictions. Without interpretability, clinicians may be hesitant to rely on AI recommendations, fearing potential errors or biases in the model. Additionally, the lack of explainability complicates the process of model improvement, as it is difficult to identify areas where the model may be failing or making incorrect predictions. Thus, the primary problem this research addresses is the need for a reliable and interpretable AI model for polyp detection that can be trusted by clinicians for use in CRC diagnosis. This study aims to bridge the gap between the high performance of AI models and the need for transparency and trust in clinical settings.

IV. RESEARCH OBJECTIVES

The overarching goal of this research is to develop an explainable AI framework for polyp detection in colorectal cancer diagnosis. Specifically, the study has three main objectives:

A. To Develop an Explainable AI Model for Polyp Detection:

The first objective is to create a deep learning model capable of accurately detecting colorectal polyps in medical images. The model will be designed with interpretability in mind, integrating techniques for visualizing and explaining its decision-making process. The goal is to ensure that the model not only performs well in terms of accuracy but also provides transparent, understandable explanations for its predictions.

B. To Evaluate the Effectiveness of Existing Explainability Techniques:

The second objective is to evaluate various explainability methods, such as Grad-CAM, LIME, and SHAP, in the context of polyp detection. These methods will be applied to the model's predictions to assess how well they highlight important features in the medical images that contribute to the diagnosis. This evaluation will help determine which techniques provide the most meaningful and actionable insights for clinicians.

C. To Compare AI-Driven Detection with Human Interpretation:

The third objective is to assess the utility of the explainable AI model by comparing its performance and interpretability with that of human clinicians. Through collaboration with medical experts, the study will evaluate how well the AI explanations align with the clinical understanding of polyp detection. This comparison will shed light on whether the AI model can serve as a reliable adjunct to human expertise and enhance the diagnostic process.

V. METHODOLOGY

The methodology for this research will consist of several key components: data collection, model design, training, and evaluation.

A. Data Collection:

The dataset used for this study will include colonoscopy images, CT colonography scans, or histopathology slides containing labeled polyps. These images will be sourced from publicly available datasets or clinical partnerships with hospitals or research institutions. The dataset will be carefully curated to ensure a diverse set of cases, representing various types and sizes of polyps, as well as varying levels of image quality.

B. Model Design and Training:

A convolutional neural network (CNN) or a variant of it, such as U-Net, will be employed for the polyp detection task. The model will be trained on the collected dataset using standard techniques, such as data augmentation and transfer learning, to improve generalizability. Hyperparameter tuning and cross-validation will be used to optimize the model's performance. During the training process, the model will learn to identify patterns in the medical images that correlate with the presence of polyps.

C. Explainability Techniques:

To make the model's decisions interpretable, we will integrate explainability methods like Grad-CAM, SHAP, and LIME. These techniques will generate heatmaps, feature importance scores, and other visualizations to illustrate the areas of the image that influenced the model's prediction. The effectiveness of these methods will be evaluated in terms of clarity and usefulness for clinicians.

D. Evaluation:

Model performance will be evaluated using standard metrics such as accuracy, sensitivity, specificity, and F1 score. Additionally, the quality of the explanations generated by the explainability techniques will be assessed through user studies involving medical professionals. Clinicians will be asked to evaluate whether the explanations help them better understand the model's decisions and how these explanations impact their diagnostic confidence.

VI. RESULTS

This section will present the outcomes of the model's performance in terms of its ability to accurately detect polyps in medical images. Key performance metrics will include accuracy, sensitivity (true positive rate), specificity (true negative rate), and precision, along with a discussion of how these metrics compare to existing state-of-the-art models in polyp detection. Additionally, the effectiveness of the explainability techniques will be evaluated by measuring how well they illuminate the model's decision-making process. Visualizations such as Grad-CAM heatmaps, LIME's local explanations, and SHAP's feature importance scores will be analyzed for clarity and relevance to clinicians. A user study will be conducted to gather feedback from radiologists or gastroenterologists on the interpretability of the explanations and how they might influence their decision-making. The results will also compare AI-based detection with human expert performance, considering both diagnostic accuracy and the interpretability of AI decisions.

VII. DISCUSSION

The discussion will analyze the implications of the results in the context of medical practice. It will explore the strengths and weaknesses of the developed AI model for polyp detection, particularly in terms of its clinical utility. While the model may demonstrate high accuracy, it is essential to consider whether the explanations provided by the AI are sufficiently intuitive and useful to clinicians. This section will also address potential challenges, such as the scalability of the approach to other medical imaging tasks, the limitations of the explainability methods, and how well the AI model integrates with existing clinical workflows. The potential for AI to serve as a reliable assistant in polyp detection will be discussed, with emphasis on how explainability can build trust and foster better collaboration between AI systems and medical professionals. Finally, the discussion will highlight potential avenues for further research, such as improving explainability techniques, exploring other AI models, and expanding the framework to cover other cancer types.

VIII. CONCLUSION

In conclusion, this research aims to develop an explainable AI framework for polyp detection in colorectal cancer diagnosis, with the ultimate goal of improving clinical decision-making and patient outcomes. The study demonstrates that it is not enough for AI systems to be accurate; they must also provide transparent and interpretable predictions to be trusted by clinicians. By applying explainability techniques to a deep learning model for polyp detection, the research shows that AI can be a valuable tool in the early detection of colorectal cancer, while also ensuring that the decision-making process remains understandable. Future work will focus on refining the interpretability of AI models, exploring their applicability to other areas of medical imaging, and investigating how these models can be integrated into clinical practice to enhance the diagnostic process.

IX. REFERENCE

[1] Esteva, A., Kuprel, B., Novoa, R. A., et al. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature, 542*(7639), 115-118. https://doi.org/10.1038/nature21056

[2] K. Sundar, V. K, K. S. N and E. Manohar, "Automated Polyp Detection in Colorectal Cancer Diagnosis using Deep Learning Techniques," *2024 8th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, Kirtipur, Nepal, 2024, pp. 1726-1731, doi: 10.1109/I-SMAC61858.2024.10714685.

[3] Tajbakhsh, N., Shin, J. Y., Gurudu, S. R., et al. (2020). Artificial intelligence in healthcare: Past, present and future. *Seminars in Cancer Biology, 65*, 3-13. https://doi.org/10.1016/j.semcancer.2019.12.002

[4] Wang, J., Yang, L., & Chen, L. (2018). Polyp detection in colonoscopy images using convolutional neural networks. *Computers in Biology and Medicine, 100*, 124-132. https://doi.org/10.1016/j.compbiomed.2018.08.001

[5] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you? Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135-1144. https://doi.org/10.1145/2939672.2939778

[6] Selvaraju, R. R., Cogswell, M., Das, A., et al. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 618-626. https://doi.org/10.1109/ICCV.2017.74

[7] Gao, X., Liu, Y., & Zhang, Y. (2019). Deep learning-based detection of polyps in colonoscopy images: A review. *Medical Image Analysis, 58*, 101554. https://doi.org/10.1016/j.media.2019.101554

[8] Caruana, R., Geiger, D., & Cohn, J. (2019). Clinical machine learning: Perspectives and opportunities. *Machine Learning, 108*(3), 1-15. https://doi.org/10.1007/s10994-019-0587-9

[9] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 4765-4774. https://doi.org/10.5555/3295222.3295344

[10] Liu, Y., & Ge, Z. (2019). Polyp detection and segmentation in colonoscopy images using deep learning: A review. *Artificial Intelligence in Medicine, 98*, 46-59. https://doi.org/10.1016/j.artmed.2019.06.003

[11] Cheng, J., Zhang, X., Wang, L., et al. (2019). A survey of deep learning in medical image analysis. *Medical Image Analysis, 42*, 60-88. https://doi.org/10.1016/j.media.2017.07.005

[12] Sunil Kumar Suvvari & DR. VIMAL DEEP SAXENA. (2024). Innovative Approaches to Project Scheduling: Techniques and Tools. Innovative Research Thoughts, 10(2), 133–143. https://doi.org/10.36676/irt.v10.i2.1481

[13] Suman Chintala, Vikramrajkumar Thiyagarajan, 2023. *"Harnessing AI for Transformative Business Intelligence Strategies", ESP International Journal of Advancements in Computational Technology (ESP-IJACT)* Volume 1, Issue 3: 81-96.

[14] Sudheer Amgothu, Giridhar Kankanala, 2024. *Adoption of Source Control Systems in the Software Industry,* ESP Journal of Engineering & Technology Advancements 4(1): 122-125.

[15] Geetesh Sanodia, "*Enhancing Salesforce CRM with Artificial Intelligence*", International Journal of Artificial Intelligence Research and Development (IJAIRD), 1(1), 2023, pp. 52-61.

[16] Amrish Solanki, Kshitiz Jain, Shrikaa Jadiga, "Building a Data-Driven Culture: Empowering Organizations with Business Intelligence," International Journal of Computer Trends and Technology, 2024; 72, 2: 46-55. [Link]

[17] Naga Ramesh Palakurti, 2023. "Evolving Drug Discovery: Artificial Intelligence and Machine Learning's Impact in Pharmaceutical Research" ESP Journal of Engineering & Technology Advancements 3(3): 136-147.

[18] Nimeshkumar Patel, 2022." QUANTUM CRYPTOGRAPHY IN HEALTHCARE INFORMATION SYSTEMS: ENHANCING SECURITY IN MEDICAL DATA STORAGE AND COMMUNICATION", Journal of Emerging Technologies and Innovative Research, volume 9, issue 8, pp.g193-g202. [Link]

[19] Rajarao Tadimety Akbar Doctor, 2016." *A METHOD AND SYSTEM FOR FLICKER TESTING OF LOADS CONTROLLED BY BUILDING MANAGEMENT DEVICES", patent* Office IN, Patent number-201641009974, Application number, 201641009974, [LINK]

[20] Dixit, A., Sabnis, A., Balgude, D., Kale, S., Gada, A., Kudu, B., Mehta, K., Kasar, S., Handa, D., Mehta, R. and Kshirsagar, S., 2023. Synthesis and characterization of citric acid and itaconic acid-based two-pack polyurethane antimicrobial coatings. *Polymer Bulletin*, 80(2), pp.2187-2216.

[21] Aparna Bhat, "Comparison of Clustering Algorithms and Clustering Protocols in Heterogeneous Wireless Sensor Networks: A Survey," 2014 INTERNATIONAL JOURNAL OF SCIENTIFIC PROGRESS AND RESEARCH (IJSPR) - ISSN: 2349-4689 Volume 04-NO.1, 2014. [Link]

[22] Apurva Kumar, "Building Autonomous AI Agents based AI Infrastructure," International Journal of Computer Trends and Technology, vol. 72, no. 11, pp. 116-125, 2024. Crossref, https://doi.org/10.14445/22312803/IJCTT-V72I11P112

[23] Chandrakanth Lekkala 2022. "Integration of Real-Time Data Streaming Technologies in Hybrid Cloud Environments: Kafka, Spark, and Kubernetes", European Journal of Advances in Engineering and Technology, 2022, 9(10):38-43. [Link]

[24] Thapliyal, P. S. Bhagavathi, T. Arunan and D. D. Rao, "Realizing Zones Using UPnP," *2009 6th IEEE Consumer Communications and Networking Conference*, Las Vegas, NV, USA, 2009, pp. 1-5, doi: 10.1109/CCNC.2009.4784867.

[25] Mihir Mehta, 2024." *Evaluating the Trade-offs Between Fully Managed LLM Solutions and Customized LLM Architectures: A Comparative Study of Performance, Flexibility, and Response Quality",* International Journal of Management, IT & Engineering, volume 14, Issue 10, [Link]

[26] *Hybrid Transformation Model: A Customized Framework for the Digital-First World* - Karthik Hosavaranchi Puttaraju - IJFMR Volume 4, Issue 1, January-February 2022.

[27] Tsaliki KC. AI-driven hormonal profiling: a game-changer in polycystic ovary syndrome prevention. Int J Res Appl Sci Eng Technol (IJRASET). 2024. https://doi.org/10.22214/ijraset.2024.61001.

[28] Next-Generation Decision Support: Harnessing AI and ML within BRMS Frameworks (N. R. Palakurti , Trans.). (2023). International Journal of Creative Research In Computer Technology and Design, 5(5), 1-10. https://jrctd.in/index.php/IJRCTD/article/view/42

[29] Lekkala, Chandrakanth, AI-Driven Dynamic Resource Allocation in Cloud Computing: Predictive Models and Real-Time Optimization (February 06, 2024). J Artif Intell Mach Learn & Data Sci | Vol: 2 & Iss: 2, Available at SSRN: https://ssrn.com/abstract=4908420 or http://dx.doi.org/10.2139/ssrn.4908420