PhD Thesis abstract

Artificial intelligence applications in the domain of digestive endoscopy

PhD student: Ioanovici Andrei-Constantin

Doctoral advisor: prof. univ. dr. Dobru Daniela

Introduction

Lower gastrointestinal endoscopy is the gold standard for colorectal cancer screening and diagnosis - the third most common cancer worldwide - yet it is estimated that approximately 26% of adenomas may be missed during the procedure. AI-based systems can improve several performance indicators, such as the adenoma detection rate; however, progress depends on datasets that are sufficiently large, heterogeneous, and well-annotated. Medical data are often unstructured, and developing methods to extract and structure these data is necessary for more comprehensive patient profiling and to design tailored, effective management strategies

Objectives

The thesis investigates whether synthetic data and pseudosynthetic data (a concept newly introduced in this work) can enhance the performance and generalizability of AI models along two complementary directions - imaging and text.

The first study introduces the concept of pseudosynthetic images as an extension of traditional augmentation tailored to colonoscopy. Real colonoscopy images were subjected to controlled transformations to reproduce the variability encountered in endoscopic practice while maintaining traceability to the original source. These data were used together with real and synthetic images (generated via GAN and DDPM) to train AI models for polyp detection, with performance evaluated by external validation on a set of real images not previously seen by the models.

The second study assessed the clinical realism of pseudosynthetic and synthetic images through a questionnaire administered to gastroenterologists (specialists/consultants and residents).

The third study aimed to develop AI models for natural language processing, using real and synthetic text datasets (endoscopy reports), to compare these models and evaluate their potential to extract useful clinical information for integration into an intelligent multimodal patient-management system.

Results

Study 1 (imaging, U-Net model). In internal validation, training on real plus pseudosynthetic images outperformed single-source experiments. On the test set for the real + pseudosynthetic regimen, results were F1 0.8832 (Dice 0.8799; IoU 0.7875), whereas combinations including synthetic data ranged from F1 0.775 for real + synthetic (GAN) to F1 0.891/0.914 for real + synthetic (DDPM) at 25k/50k epochs. Using all sources, the model reached F1 0.912 (Dice 0.902; IoU 0.823). In the head-to-head comparison of the two synthetic generation methods, the GAN regimen recorded F1 0.919 (Dice 0.911) compared with the diffusion method F1 0.885 (Dice 0.867). On external validation with CVC-Clinic-DB (612 image pairs), results were: real F1 0.637 (Dice 0.582; IoU 0.495), pseudosynthetic F1 0.750 (Dice 0.743; IoU 0.646), synthetic (GAN) F1 0.494 (Dice 0.109; IoU 0.083), and DDPM 25k/50k/100k F1 0.715/0.768/0.729. For combinations: real + pseudosynthetic F1 0.780 (Dice 0.764; IoU 0.677), real + synthetic (GAN) F1 0.693, real + pseudosynthetic + synthetic (GAN) F1 0.777, real + pseudosynthetic + synthetic (DDPM 25k) F1 0.751, and real + pseudosynthetic + synthetic (GAN+DDPM) F1 0.768.

Study 2 (clinical/perceptual realism). In the gastroenterologist survey (24 images: 8 real, 8 pseudosynthetic, 8 synthetic - GAN/DDPM), overall accuracy was 61.2% (95% CI: 57.7–64.6%), with no difference between residents (62.3%) and specialists/consultants (59.8%; $p = 0.54$); Fleiss' $\kappa = 0.30$ (95% CI: 0.15–0.43). By category, sensitivity/precision were: real 70.7%/62.2%; pseudosynthetic 51.6%/58.9%; synthetic 61.3%/62.1%. GAN images were recognized as synthetic in 100% of evaluations (128/128; 95% CI 97.1–100%), whereas for DDPM recognition as synthetic was 22.7% ($p < 0.001$).

Study 3 (text, Romanian NER). On the real set, ModelR (trained on real reports) achieved F1 0.8978, ModelM (mixed real + synthetic, 1:1) F1 0.8949, and ModelS (synthetic-only) F1 0.3326; on the synthetic set, ModelS/ModelM reached F1 0.9922, while ModelR had F1 0.4686. On the global evaluation (combined real + synthetic), ModelM recorded F1 0.9425 (Precision 0.9491; Sensitivity 0.9348), outperforming ModelR (F1 0.6976) and ModelS (F1 0.6404).

Discussion and Conclusions

This thesis introduces, as a first, the concept of pseudosynthetic data, defined as real colonoscopic images subjected to controlled augmentation that simulates examination variability as if originating from multiple colonoscopies of the same patient. This approach provides an ethical, feasible way to enrich datasets while preserving traceability and clinical plausibility.

Overall, the findings converge on the same conclusion: a mixed-source strategy (real, synthetic, pseudosynthetic) is effective both for visual tasks (U-Net segmentation) and for clinical NLP (NER). By narrowing the gap between technical innovation and clinical impact - from fewer missed lesions and improved diagnosis to more effective screening and personalized management - this thesis outlines a realistic pathway to implementation.