A Hypothesis-Driven Framework for Automatic Gastrointestinal Diagnosis from Ultrasound in Austere Settings

Machine Learning and Artificial Intelligence in Military Health, Readiness and Performance

Juan P. Wachs
James H. and Barbara H. Greene Professor
Edwardson School of Industrial Engineering
Weldon School of Biomedical Engineering
Purdue University, West Lafayette, IN, USA
School of Medicine, Indiana University, Indianapolis, IN, USA

PURDUE UNIVERSITY

SCHOOL OF MEDICINE
INDIANA UNIVERSITY

Disclosure

Juan Wachs, PhD has disclosed no relevant, real or apparent personal or professional financial relationship with proprietary entities that produce healthcare goods and services.

Introduction

The clinical gap

- Austere and military settings often lack gastrointestinal specialists, delaying diagnosis and treatment decisions.
- Ulcerative Colitis requires timely, repeated monitoring to prevent irreversible bowel damage, colorectal cancer, and emergency surgery.

The technical gap

- Intestinal ultrasound is portable and non-invasive, but its use is limited by expert dependence and interpretation variability.
- Standard AI models need large annotated datasets and lack interpretability.

In this study, we introduce REFINE: a hypothesis-driven generative AI framework with logic and context-based reasoning, for accurate, data-efficient and interpretable diagnosis for gastrointestinal decision-making in austere environments. We test our method on B-Mode ultrasound of Ulcerative Colitis patients, for colon wall stratification status.
REFINE gives 87% accuracy, exceeding baseline VLMs and other AI models for the task.

Methodology:
Incorporating Clinical Knowledge into AI Reasoning

Use a pretrained vision language model (VLM) to combined experimental setup and clinical insights into a unified textual prompt – generates initial hypotheses.

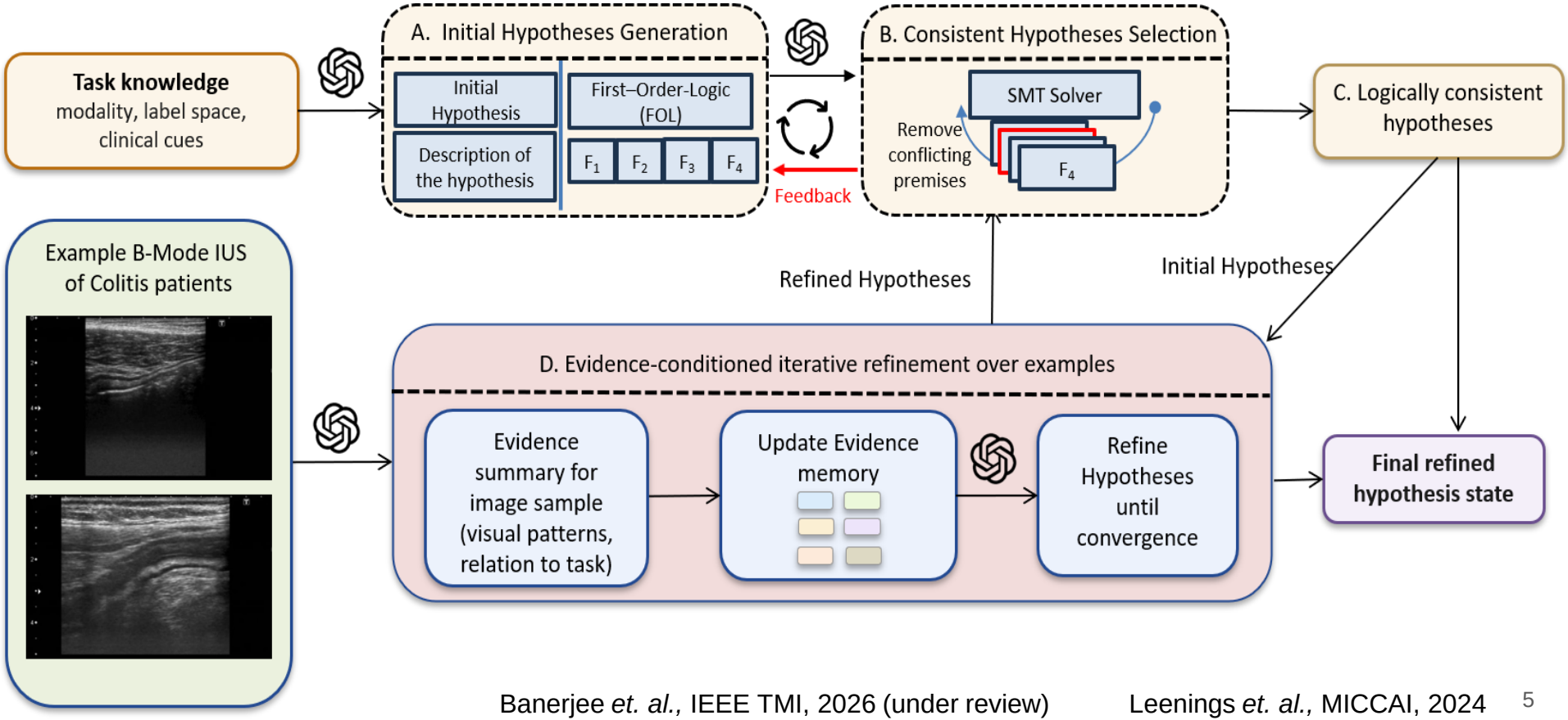
System Prompt: You are a medical professional interpreting a B-Mode ultrasound image of the colon for Ulcerative Colitis patients....

Hypothesis: ... Preserved stratification: Intermittent banding remains visible along part of the bowel wall...Lost: ..no separable layer boundaries are visible ...

Rahman *et. al,* MIUA, 2026

Banerjee *et. al.,* IEEE TMI, 2026 (under review)

Methodology: REFINE Framework

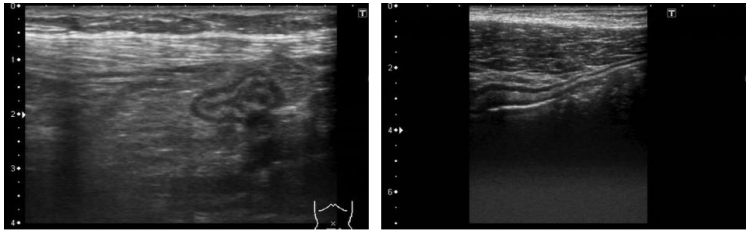


Banerjee *et. al.,* IEEE TMI, 2026 (under review)

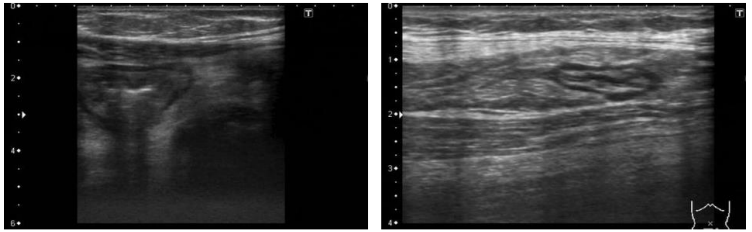
Leenings *et. al.,* MICCAI, 2024

Experiments: Ulcerative Colitis Ultrasound Dataset

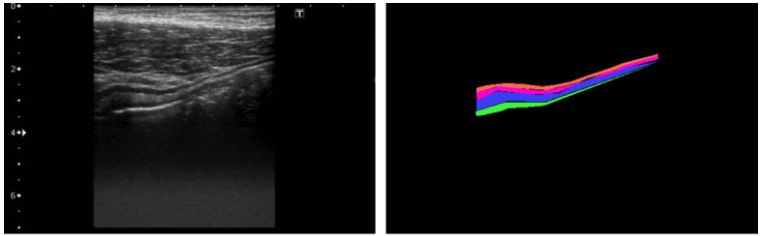
- Publicly available C-TRUS dataset
- Dataset contains 800 B-Mode Ultrasound images collected from 13 UC patients
- 275 good images quality selected for analysis
- Dataset was annotated for the following:
    - 4-layer colon wall stratification status (preserved (P) or lost (L))
    - Segmentation of the 4-layers of the colon wall for class for preserved (P) class images



Example of Preserved (P) colon wall stratification



Example of Lost (L) colon wall stratification



4 layer stratification – mucosa (green), submucosa (blue), muscularis (pink), serosa (orange)

Leenings *et. al.,* MICCAI, 2024

Experiments: Models and Evaluation Setup

❑ **Task**:

Colon Wall Stratification Status

Binary Classification

Classes:

- ➢ Preserved (P)
- ➢ Lost (L)

❑ **Reasoning Variants** 🧠

- ➢ Hypothesis Only
- ➢ Hypothesis + Logical Consistency
- ➢ Hypothesis + Logical Consistency

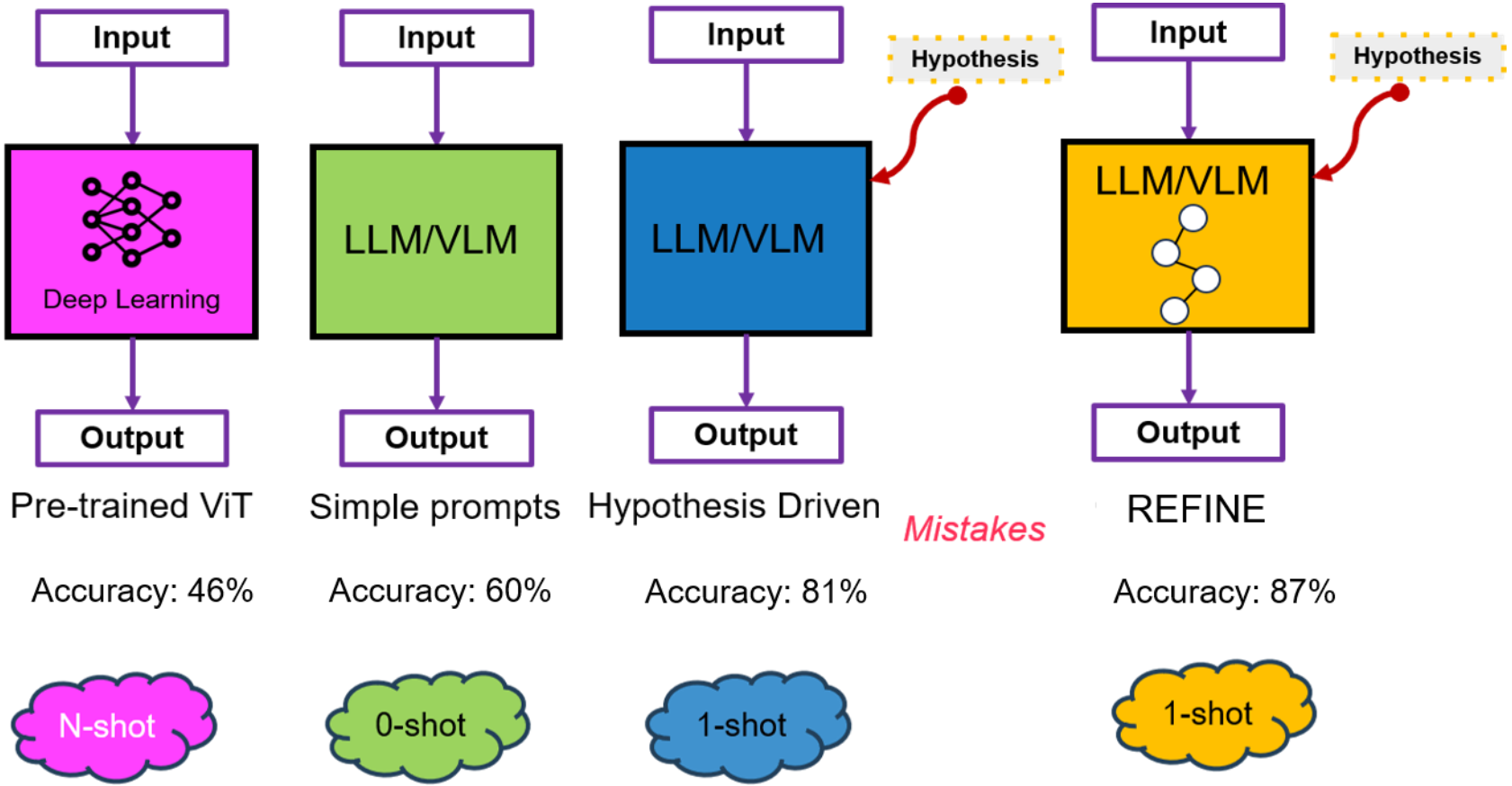
+ Visual consistency (REFINE)

❑ **VLM Models**

- ➢ GPT-4o, GPT-4o-mini, GPT-4-Turbo,
- ➢ Gemini-2.0, Gemini-1.5
- ➢ Qwen-2-VL-2B

REFINE Improves Diagnostic Accuracy

Demonstrated on GPT4o

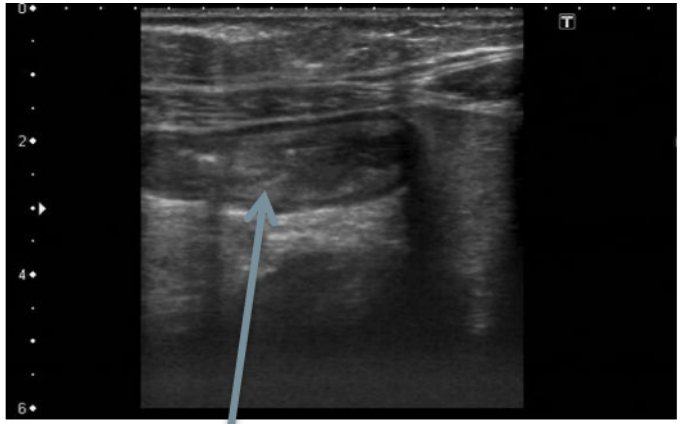


Performance Across Multiple AI Models

Model	Accuracy	F1	Precision	Recall
GPT-4o (Base)	0.60	0.66	0.69	0.63
GPT-4o + REFINE	**0.87**	0.78	0.83	0.74
GPT-4o-mini (Base)	0.61	0.67	0.86	0.59
GPT-4o-mini+REFINE	0.81	**0.84**	**0.88**	0.81
Gemini-2.5 (Base)	0.33	0.18	0.78	0.10
Gemini-2.5+REFINE	0.71	0.78	0.86	0.72
Qwen2-VL-2B (Base)	0.66	0.77	0.76	0.79
Qwen2-VL-2B+REFINE	0.74	**0.84**	0.75	**0.96**
CLIP-Fewshot	0.46	0.45	0.38	0.42
ResNet-34-Fewshot	0.12	0.15	0.13	0.12

REFINE Improves AI Reasoning

Input: B-Mode Ultrasound image



Ground Truth Stratification Status: Lost (L)

The bowel wall appears diffusely hypoechoic and homogeneous, with no separable alternating layers or consistent banding along the segment, supporting lost stratification.
Prediction: Lost (L)

GPT4o+Hypotheses refinement

Correct classification

Residual layered bands remain visible, though fuzzy, partially merged, and intermittently disrupted; therefore stratification is mostly preserved rather than completely lost.
Prediction: Preserved (P)

GPT4o

Incorrect classification

Banerjee *et. al.,* IEEE TMI, 2026 (under review)

Leenings *et. al.,* MICCAI, 2024

Takeaways

General-purpose VLMs struggle with specialized medical data, underscoring the need for domain-guided reasoning to ensure accurate and reliable clinical predictions.

- GPT-4o misclassifies a lost stratification for blurred by dismissing the visual cue supporting the relevant hypotheses.
- GPT-4o + REFINE correctly classifies the stratification as lost by maintaining consistency between the visual observation and the generated hypotheses.
- Demonstrates how REFINE enhances diagnostic accuracy and interpretability through explanation-guided reasoning and consistency checking.

Acknowledgements

The AI model development was supported by the Office of the Assistant Secretary of Defense for Health Affairs under Award No. W81XWH-21-2-0030.

We would like to thank Pronoma Banerjee, Masudur Rahman (Purdue University) and Sanjay Mohanty (Indiana University School of Medicine) for their contribution to the research. We would also like to thank Anuva Shah (Sunset High School, Portland, Oregon), Jason Wu (Langley High School. Virginia) and Zackary Weinreb (Purdue University) for their help.

References

1. R. Leenings, M. Konowski, N. Winter, J. Ernsting, L. Fisch, C. Barkhau, U. Dannlowski, A. Lügering, Andreas, X. Jiang (2024). C-trus: A novel dataset and initial benchmark for colon wall segmentation in transabdominal ultrasound. In International Workshop on Advances in Simplifying Medical Ultrasound (pp. 101-111), MICCAI. Cham: Springer Nature Switzerland.
2. P Banerjee, M. M. Rahman, A. Shah, S. Mohanty, F. Zor and J. Wachs, “Domain-specific hypothesis generation andrefinement for few-shot medical imageunderstanding with Vision Language Models” in IEEE Transactions in Medical Imaging (Manuscript Under Review).
3. M. M. Rahman, M. El Masry, G. Gordillo, and J. P. Wachs, “Knowledge-driven hypothesis generation for burn diagnosis from ultrasound withvision-language model,” in Medical Image Understanding and Analysis. Springer Nature Switzerland, 2026, pp. 275–288
4. M. M. Rahman, M. El Masry, G. Gordillo, and J. Wachs, “Automatedunified reasoning with vision-language models for multi-modal burnassessment,” in Proceedings of the AAAI Conference on ArtificialIntelligence, vol. 40, no. 47, 2026, pp. 40 402–40 408.

For questions, please email:
Juan Wachs (jpwachs@purdue.edu) or
Pronoma Banerjee (baner102@purdue.edu)