Towards Robust, Reliable, and Generalized Medical AI

Chenyu You

Department of Applied Mathematics & Statistics
Department of Computer Science
Stony Brook University
chenyu.you@stonybrook.edu

Recent advances in AI have led to remarkable progress in biomedical image analysis, yet reliable clinical deployment remains challenging. Three issues are particularly critical: 1) limited labeled data, which restricts the training of high-capacity models; 2) distribution shifts across heterogeneous clinical environments, which weaken robustness and generalization; and 3) the absence of theoretical guarantees, which hinders safety and trustworthiness. My research addresses these challenges by developing label-efficient and class-imbalance learning methods, unifying algorithm design with theoretical analysis, and advancing medical foundation models capable of generalizing across diverse populations and modalities.

Part 1: Data-Efficient Representation Learning for Medical Imaging A major bottleneck in medical AI is the scarcity of labeled data and the imbalance of disease categories, both of which hinder robust model training. My work addresses these challenges through the MONA framework (You et al. 2024a), which unifies two key contributions: ***(i)*** learning effective anatomical representations from extremely limited labeled data by leveraging self-supervised training strategies on 3D medical images, and ***(ii)*** improving robustness against long-tailed distributions by dynamically balancing representations across underrepresented classes. From these perspectives, MONA enables the models to be robust and generalizable even under extremely limited label settings, moving medical AI a step closer to deployment across diverse clinical settings.

Part 2: Reliable Algorithms with Theoretical Guarantees. Safety-critical domains like healthcare demand models with both empirical accuracy and certifiable reliability. My work (You et al. 2023) integrates learning-theoretic principles into medical AI, such as variance-reduction sampling and class-divergence analysis, yielding robust objectives with provable guarantees. I also investigated visual interpretability, showing how attention mechanisms and feature alignment uncover the anatomical reasoning of deep models (You et al. 2022). These contributions provide a unified framework that couples theoretical understanding with clinical relevance.

Part 3: Towards Trustworthy Medical Foundation Models. Medical foundation models (MedFMs) represent a transformative step toward universal clinical AI systems. I have contributed to build the *first* MedFMs (Ma et al. 2024) capable of generalization across modalities and institutions, while also addressing their limitations. For example, I developed the robust calibration method (You et al. 2024b) that improve group robustness under subpopulation shifts. These contributions advance MedFMs from powerful prototypes to trustworthy clinical tools by improving their adaptability, efficiency, and trustworthy across diverse patient populations.

Future Directions. I aim to broaden my research in three key directions: 1) develop self-supervised and knowledge-augmented methods for robust performance under scarce annotations and heterogeneous data (You et al. 2024a); 2) advance learning theory for medical AI through sample complexity analysis, statistical guarantees, and causality-based robustness to distribution shifts (You et al. 2025); 3) design interactive clinical AI agents that continuously learn from workflows and integrate multimodal knowledge, ensuring adaptability and reliability in deployment. Ultimately, my goal is to build robust, reliable, and generalizable biomedical AI that are clinically actionable.

References

Ma, J.; He, Y.; Li, F.; Han, L.; You, C.; and Wang, B. 2024. Segment Anything in Medical Images. *Nat. Commun.*

You, C.; Dai, H.; Min, Y.; Sekhon, J.; Joshi, S.; and Duncan, J. S. 2025. Uncovering Memorization Effect in the Presence of Spurious Correlations. *Nat. Commun.*

You, C.; Dai, W.; Liu, F.; Min, Y.; Dvornek, N. C.; Li, X.; Clifton, D. A.; Staib, L.; and Duncan, J. S. 2024a. Mine yOur owN Anatomy: Revisiting medical image segmentation with extremely limited labels. *IEEE TPAMI*.

You, C.; Dai, W.; Min, Y.; Liu, F.; Clifton, D. A.; Zhou, S. K.; Staib, L.; and Duncan, J. S. 2023. Rethinking semi-supervised medical image segmentation: A variance-reduction perspective. *NeurIPS*.

You, C.; Min, Y.; Dai, W.; Sekhon, J.; Staib, L.; and Duncan, J. S. 2024b. Calibrating Multi-modal Representations: A Pursuit of Group Robustness Without Annotations. In *CVPR*.

You, C.; Zhao, R.; Liu, F.; Dong, S.; Chinchali, S.; Topcu, U.; Staib, L.; and Duncan, J. S. 2022. Class-aware adversarial transformers for medical image segmentation. *NeurIPS*.