Enhancing Trustworthiness in Medical Imaging through Interpretability and Automated Bias Detection

Deep learning (DL) has achieved remarkable performance across medical imaging tasks – ranging from disease classification from chest X-rays (CXR) to breast cancer risk prediction, yet its clinical adoption remains limited. Two critical gaps constrain its trustworthiness: **1. Lack of Interpretability.** Despite high performance, current "black-box" DL models fail to provide transparency in their decision-making process. Clinicians require interpretability to ensure predictions align with clinical reasoning (e.g., using the cardiothoracic ratio to identify cardiomegaly). Predictions without clear explanations impede trust and practical integration into clinical workflows. **2. Susceptibility to Shortcuts and Biases.** DL models in healthcare often rely on spurious correlations, leading to inconsistent performance across demographic groups and perpetuating health disparities. Existing bias detection methods lack the necessary domain knowledge, rely heavily on expensive annotations, and fail to identify hidden biases systematically. Similarly, current mitigation strategies address single biases, requiring prior knowledge and annotations. This thesis tackles these challenges by designing advanced AI algorithms and demonstrating their effectiveness in real-world healthcare applications.

To address these gaps, this thesis proposes a three-aim study to develop (1) a post-hoc interpretable model using First-Order Logic (FOL) rules to explain the blackbox, (2) an automated debugger that explains the errors of the blackbox via language, and (3) apply this debugger to the domain of breast cancer risk prediction, focusing on bias detection and mitigation. Aim 1 attempts to explain the blackbox via concept-based rules. We hypothesize that these concepts are invariant across domains. Standard transfer learning requires large training data across domains. So, as a part of Aim 1, we leverage the learned concepts from the interpretable model to perform data-efficient transfer learning. However, in Aim 1, we face the challenge of concept annotation since explaining the entire blackbox requires labeling high-level concepts. However, fully parsing the entire blackbox can be unnecessary if we only need to clarify error modes; thus, Aim 2 targets blackbox misclassification, specifically, using language-based auditing to detect and mitigate errors. Finally, Aim 3 applies our automated debugging approach to breast cancer risk prediction. No prior work has systematically evaluated biases in state-of-the-art models (e.g., MIRAI) across demographic subgroups, and no dedicated vision-language model (VLM) exists for fine-grained mammogram-report reasoning. So, in Aim 3, we will develop the first VLM for the breast cancer domain. Overall, this thesis aims to enhance reliability and equity in clinical AI systems. Specifically, this thesis will:

Aim 1: Develop an Interpretable DL Framework for CXR and Enable Data-Efficient Transfer learning

1.1. Extract a mixture of interpretable models from the blackbox. We will first introduce an iterative method that carves out a mixture of interpretable models from a pretrained blackbox. Each interpretable model focuses on a specific subgroup of samples. It explains predictions by generating First-Order Logic (FOL) rules, which are constructed from the black box's underlying learned concepts. This ensures the reasoning aligns with clinical standards and enhances transparency.

1.2. Leverage FOL-extracted rules for data-efficient transfer learning. Since FOL rules remain invariant across domains, we will develop a transfer learning algorithm to adapt interpretable models trained on the MIMIC-CXR dataset to the Stanford-CXR dataset, leveraging limited training data without any concept annotations in the Stanford-CXR dataset.

Aim 2: Build an Automated Debugging Approach for Black-Box Errors and Misclassifications

2.1. Explain black-box errors via language We will develop a debugging algorithm that uses LLM and VLM to parse the captions (or radiology reports), model outputs, and associated metadata to identify biased subgroups. By translating model mistakes into natural language hypotheses via LLM (e.g., "misclassification might be due to a spurious label correlation with chest tubes"), our system will systematically identify and explain the failure modes of the blackbox, offering actionable insights into its error patterns. For mitigation, this method generates pseudo-labels for the biased subgroups corresponding to each hypothesis and fine-tune the linear head of the blackbox.

Aim 3: Apply the Debugger to Breast Cancer Risk Prediction and Enhance Vision-Language Reasoning

3.1. Systematic bias assessment of SOTA breast cancer risk predictors. We will conduct a rigorous study of the state-of-the-art (SOTA) breast cancer risk predictor – MIRAI's performance across critical patient attributes (age, breast density, BI-RADS, and race), quantifying disparities and identifying subpopulations at elevated risk of misclassification. This analysis will provide a foundational understanding of existing biases and set the stage for targeted interventions.

3.2. Develop a dedicated mammography VLM with visual instruction tuning. We will develop the first VLM trained on mammogram–report pairs, ensuring it captures domain-specific nuances (e.g., microcalcifications, breast density). Then, through visual instruction tuning, we will enhance the reasoning capability of the VLM. This specialized VLM forms the backbone for debugging (from Aim 2) in mammography—enabling the automatic generation of clinically relevant hypotheses about misclassifications and guiding corrective strategies that mitigate biases uncovered in 3.1.

Impact: By seamlessly integrating interpretability (Aim 1) and debugging (Aim 2), and then applying these capabilities to a high-stakes domain (Aim 3), this work will deliver practical tools for building trustworthy and generalizable DL models in medical imaging. Collectively, these aims will improve fairness, transparency, and reliability in clinical AI systems, helping ensure that cutting-edge deep learning models truly benefit diverse patient populations.