Bringing Interpretability and Medical Explainability to Deep Learning: Diagnosis of Age-Related Macular Degeneration

Shi, Lily (School: The Harker School)

The inability of deep learning (DL) models to explain their output, known as the AI “Black Box” problem, greatly hinders clinical adoption of DL models in healthcare. This study is the first to identify interpretability and medical explainability as two distinct challenges of the “Black Box” and propose an innovative approach to address both challenges in an end-to-end DL model. Interpretability refers to the model’s ability to reveal regions of the image used in its decision process; medical explainability refers to the model’s ability to explain results in medical terms understandable to doctors and patients. Implemented through the diagnosis of age-related macular degeneration (AMD) based on fundus images from the ADAM dataset, this model generates four outputs: (1) AMD classification; (2) a heat map revealing regions of image used in decision-making (adding interpretability); (3) pixel-level segmentation of AMD-related biomarkers to explain the diagnosis (adding medical explainability); (4) HSS, a novel metric created in this study, to measure the spatial overlap between heat map regions and AMD lesion segmentation (quantifying medical relevance of heat map regions). The model achieves: 0.96 AUC, 0.95 accuracy, 0.97 sensitivity, 0.89 specificity for AMD classification, and extracts pertinent biomarkers (drusen, exudate, hemorrhage, scar) in segmentation with dice similarity coefficient from 0.34 to 0.76. An average HSS of 0.69 suggests substantial medical relevance of heat map regions. By enhancing interpretability and medical explainability, my study contributes significantly to building trust in and improving acceptance of AI in healthcare.