An Adaptive Grad-CAM Thresholding Technique to Improve AI Interpretability in MRI-Based Prostate Cancer Risk Stratification

Sreepathi, Yash (School: Broad Run High School)

Artificial Intelligence (AI) models like Convolutional Neural Networks (CNNs) are becoming powerful tools in medical imaging, but their “black box” nature limits clinical use. Doctors need to understand why an AI made a decision. This project aimed to improve the interpretability of CNN-based cancer predictions by refining Gradient-weighted Class Activation Maps (Grad-CAM) using an adaptive thresholding technique. A modified DenseNet121 CNN was trained using prostate MRI data from the ProstateX dataset to classify scans as Clinically Significant (CS) or Non-Clinically Significant (NCS). Grad-CAM heatmaps were then generated and refined using two methods: global thresholding, which applies a fixed cutoff across the entire image, and adaptive thresholding, which uses a sliding window to adjust thresholds based on local image regions. Clinical interpretability of thresholded heatmaps was assessed by comparing the activation areas between CS and NCS MRI scans. While global thresholding produced vague and overlapping heatmaps, adaptive thresholding revealed a clear difference between CS and NCS cases. Three adaptive parameter combinations yielded statistically significant separation, with up to a 93.6% difference in activation area between CS and NCS. Because medical MRI scans can vary greatly from case to case, this adaptive technique better accounted for those differences while still producing consistent, quantitative separation between CS and NCS regions. These findings suggest that adaptive thresholding improves the clarity and precision of Grad-CAM heatmaps, offering a more clinically useful explanation of AI decisions. This method could support greater trust in medical AI tools and help enable their implementation in real-world cancer diagnosis.

Awards Won: