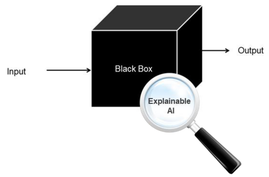
Explainable AI for Medical Image Classification and Segmentation

Abu Noman Md Sakib
The University of Texas at San Antonio

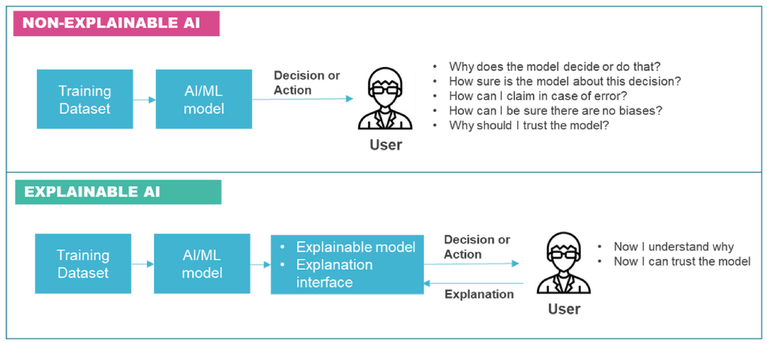
INTRODUCTION

Artificial Intelligence (AI) has significantly improved the accuracy of disease diagnosis and medical image segmentation. However, deep learning models often function as *black boxes*, making their decision-making process difficult to interpret. This lack of transparency hinders medical professionals' trust in AI-driven diagnostic tools.



Why XAI

- AI models in healthcare require transparency for safe adoption.
- Trust and interpretability are essential for medical professionals.
- Ethical concerns necessitate accountability and fairness.

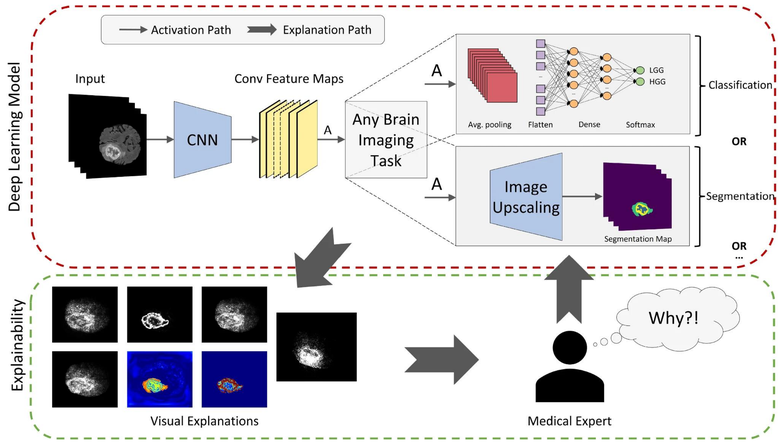


This research focuses on integrating Explainable AI (XAI) techniques to enhance the interpretability of deep learning models used in medical imaging. By visualizing critical features influencing AI predictions, we aim to make medical AI systems more reliable and acceptable for clinical applications.

OBJECTIVES

- Develop an explainable AI framework for medical imaging
- Achieve high accuracy in classifying diseases from images
- Perform precise segmentation of anatomical or pathological regions
- Increase clinician trust through transparent AI decisions
- Identify and visualize key features driving predictions
- Validate model explanations against expert clinical annotations
- Ensure compatibility with real-world clinical workflows

METHODOLOGY



Dataset and Preprocessing

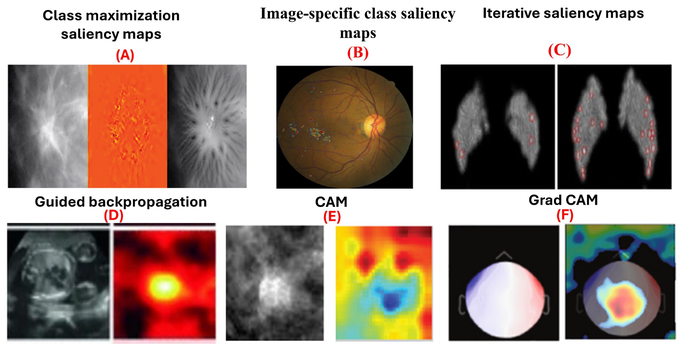
- Utilizing medical imaging datasets (CT/MRI scans).
- Preprocessing steps include normalization, augmentation, and segmentation.

Deep Learning Models

- Implementation of Convolutional Neural Networks (CNNs) for classification.
- Application of U-Net and Transformer-based models for segmentation.

XAI Techniques

- **Activation-Based Methods:** Grad-CAM, Grad-CAM++, and Score-CAM to generate heatmaps highlighting critical image regions.
- **Feature Attribution Methods:** Guided backpropagation, Layer-wise Relevance Propagation (LRP) to analyze model decisions. Shapley values to measure feature importance.

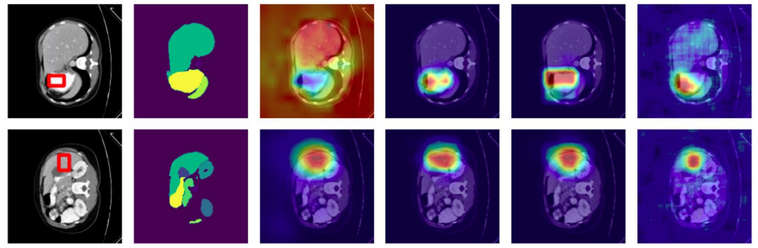


Evaluation Metrics

Segmentation Performance: Dice coefficient, Intersection over Union (IoU).
Classification Accuracy: Precision, Recall, F1-score, Accuracy, Sensitivity, Specificity.
XAI Effectiveness: Case studies evaluating interpretability improvements

RESULTS

- Heatmaps generated by Grad-CAM and Score-CAM effectively highlight important regions for classification and segmentation.
- LRP and Shapley values provide deeper insights into model decision-making by identifying key influential features.
- XAI methods improve the interpretability of AI models, making their decisions more understandable to healthcare practitioners.
- Benchmarking results show that incorporating XAI does not significantly compromise model performance while enhancing explainability.



Qualitative Results

- Trust Score: Clinician confidence rating
- Feature Concordance: AI vs. expert agreement
- Satisfaction Index: Interpretability feedback

Quantitative Results

- Accuracy: Correct classification rate
- AUC-ROC: Classification performance
- Dice: Segmentation overlap
- IoU: Segmentation precision
- Inference Time: Processing speed

CONCLUSIONS

- **Impact on Healthcare**: Increased transparency fosters trust in AI-assisted diagnosis.
- **Ethical AI Adoption**: Promotes accountability and fairness in medical applications.
- **Future Directions**:
  - Integrating XAI with real-time AI decision-support tools in clinical settings.
  - Exploring advanced XAI techniques for multimodal medical imaging.

REFERENCES

R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh and D. Batra, "Grad CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017, pp. 618-626

Aditya Chattopadhay et al. "Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks". In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). 2018, pp. 839–847.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. MICCAI, 234–241.

Hasany, Syed Nouman, Caroline Petitjean, and Fabrice Mériaudeau. "Seg-xres cam: Explaining spatially local regions in image segmentation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

Petsiuk, Vitali, Abir Das, and Kate Saenko. "Rise: Randomized input sampling for explanation of black-box models."