Explainable Artificial Intelligence Methods on Medical Image Classification and Segmentation

Abu Noman Md Sakib
The University of Texas at San Antonio, 2025

Supervising Professor: Maryam Tabar, Ph.D.

The advancement of Artificial Intelligence (AI) in medical imaging has significantly improved the accuracy of disease diagnosis and treatment planning. However, deep learning models, particularly in medical image classification and segmentation, are often criticized for their lack of transparency and interpretability. The black-box nature of these models raises concerns about reliability, and the potential risks associated with incorrect predictions in high-stakes medical applications. This challenge makes it difficult for medical professionals to trust and effectively utilize deep learning models. The primary goal is to explore how Explainable AI (XAI) techniques can enhance trustworthiness and interpretability in deep learning-based medical imaging models. Understanding how these models conclude their predictions is crucial for gaining acceptance among healthcare practitioners and ensuring reliable healthcare applications. This research is arguably significant because of its potential to bridge the gap between AI advancements and healthcare usage. With increasing reliance on AI models in medical imaging, enhancing their interpretability can lead to better diagnostic support, and improved patient outcomes. Exploring XAI in the medical domain aligns with ethical AI principles by promoting transparency and accountability. This research will include the application of activation based visual methods such as Grad-CAM, Grad-CAM++, and Score-CAM to generate heatmaps that highlight critical regions contributing to classification and segmentation tasks that helps understand more about the visual outputs. These heatmaps will separate the specific portions from inputs to distinguish between the parts that is helping the most in prediction task. Also, this research will include feature attribution methods similar to Layer-wise Relevance Propagation (LRP), and Shapley values to get a deeper understanding of which features influence predictions. The effectiveness will be analyzed by reviewing

case studies and benchmarking XAI approaches. The result analysis will be shown using the latest work that integrated XAI methods with medical image segmentation. The findings will be combined to highlight best practices, challenges, and future research directions in the field. This research will also provide insights into how XAI can improve medical image analysis while maintaining high classification and segmentation performance. This research will contribute to the broader discussion on the role of XAI in medical imaging and its potential to improve trust and usability in clinical AI applications.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

Artificial Intelligence (AI) has revolutionized medical imaging by enhancing diagnostic accuracy and improving treatment planning. Deep learning models, particularly convolutional neural networks (CNNs), have demonstrated remarkable success in tasks such as disease classification, lesion detection, and image segmentation (Roby et al., 2025; Ronneberger, Fischer, & Brox, 2015; Zhou et al., 2020). For example, deep learning-based models have achieved expert-level performance in detecting diabetic retinopathy, with sensitivity and specificity exceeding 90% in some studies (Mahmood, Rehman, Saba, Nadeem, & Bahaj, 2023). Similarly, AI-assisted radiology systems have been shown to reduce diagnostic errors by up to 30% in mammography screening (Farrag, Gad, Fadlullah, Fouda, & Alsabaan, 2023). Despite these advancements, a critical challenge remains: the lack of transparency and interpretability in deep learning models.

Deep learning models are often referred to as "black boxes" due to their complex and opaque decision-making processes. This opacity raises concerns in high-stakes medical applications where incorrect predictions can lead to misdiagnoses and adverse patient outcomes. Studies indicate that over 60% of radiologists express concerns regarding the interpretability of AI-based diagnostic tools, and regulatory bodies such as the U.S. Food and Drug Administration (FDA) emphasize the need for explainability in AI-driven healthcare systems (Holzinger, 2018). The limited interpretability of these models hinders trust and adoption among healthcare professionals, creating a significant barrier to widespread clinical integration (Chen, Gomez, Huang, & Unberath, 2022).

Efforts to address AI interpretability in medical imaging have led to the emergence of Explainable AI (XAI) techniques. These techniques aim to make deep learning models more transparent by providing human-interpretable explanations of their decisions. Existing XAI methods can be broadly categorized into visual explanation techniques and feature attribution methods (Arrieta et al., 2020). Techniques such as Gradient-weighted Class Activation Mapping (Grad-CAM) (Selvaraju et al., 2017) and its variants (Grad-CAM++, Score-CAM) (Chattopadhay, Sarkar, Howlader, & Balasubramanian, 2018; Wang et al., 2020) generate heatmaps to highlight the most relevant

image regions for a model's decision. Local Interpretable Model-agnostic Explanations (LIME) (Ribeiro, Singh, & Guestrin, 2016) evaluate how consistently a model makes predictions by slightly changing the input images and observing how the model's output changes. Additionally, feature attribution methods like LRP (Bach et al., 2015) and Shapley values (Lundberg & Lee, 2017) provide insights into how individual features contribute to predictions. In addition, counterfactual explanations (Wachter, Mittelstadt, & Russell, 2017) explore alternative outcomes, ensuring reliability and regulatory compliance. While these methods have shown promise, there is still a gap in systematically evaluating their effectiveness for medical image segmentation and classification tasks. Furthermore, the integration of XAI techniques into clinical workflows remains an open challenge, requiring further research to improve their usability and reliability.

This research aims to investigate the role of XAI in enhancing the interpretability and trustworthiness of deep learning models used in medical image classification and segmentation. Specifically, the study will explore the application of activation-based visual methods and feature attribution techniques to explain model decisions. By analyzing and benchmarking XAI techniques, the research seeks to bridge the gap between AI advancements and their practical adoption in healthcare. The research will focus on evaluating and implementing various XAI methods to improve interpretability in medical imaging applications. Key components of the methodology include:

- Application of Activation-Based Visual Methods such as Grad-CAM, Grad-CAM++, and Score-CAM will be applied to highlight critical image regions that influence classification and segmentation decisions.

- Feature Attribution Analysis Techniques such as LRP and Shapley values will be used to assess the contribution of individual features to model predictions.

- The effectiveness of XAI methods will be analyzed through case studies using publicly available medical image segmentation datasets.

- Quantitative and qualitative assessments will be conducted to evaluate the reliability, consistency, and usability of XAI explanations in clinical settings.

The intellectual merit of this research lies in its systematic evaluation of XAI techniques for medical image analysis, providing insights into their strengths, limitations, and practical applicability. The findings will contribute to the growing field of trustworthy AI by demonstrating how explainability can enhance confidence in AI-driven diagnostics. From a broader impact perspective, the study will help advance ethical AI adoption in healthcare by promoting transparency and accountability. By improving the interpretability of deep learning models, the research will facilitate better collaboration between AI systems and medical professionals, ultimately leading to more informed clinical decision-making and improved patient outcomes. The insights gained from this study could also inform regulatory guidelines and best practices for integrating AI into medical imaging workflows. This research aligns with the broader goal of making AI-driven healthcare applications more trustworthy and accessible, paving the way for future advancements in explainable medical imaging technologies.

CHAPTER 2: LITERATURE REVIEW

To structure the bibliography on enhancing trustworthiness and interpretability in deep learning-based medical imaging models through XAI techniques, the related papers can be categorized into two primary themes:

2.1 Deep Learning Applications in Medical Image Analysis

This category includes research that applies deep learning algorithms to medical image analysis tasks such as classification, segmentation, and diagnosis. U-Net introduced a convolutional neural network architecture tailored for biomedical image segmentation. Its symmetric encoder-decoder structure with skip connections enables precise localization, making it highly effective even with limited training data. U-Net has become a foundational model in medical image segmentation tasks (Ronneberger et al., 2015). M2Net presents a multi-modal, multi-channel network designed to predict overall survival time in brain tumor patients. By integrating various imaging modalities, the model captures comprehensive features, enhancing prognostic accuracy. This approach underscores the potential of combining diverse data sources in medical predictions (Zhou et al., 2020).

SESV introduced a segmentation method that predicts and corrects errors in medical images. Leveraging deep convolutional neural networks, it enhances segmentation accuracy by iteratively refining predictions. This technique demonstrates improved performance in complex medical imaging scenarios (Xie, Zhang, Lu, Shen, & Xia, 2021). AAAUnet proposed a patch-based dilated U-Net model for automatic segmentation of abdominal aortic aneurysms from computed tomography angiography. By incorporating dilation and patch-based processing, it achieves high accuracy in delineating aneurysm regions, facilitating better clinical assessments (Roby et al., 2025). Another study presented an explainable AI system for mammogram tumor segmentation that preserves local resolution. By maintaining spatial details, the system enhances the interpretability of segmentation results, aiding clinicians in accurate tumor identification and analysis (Farrag et al.,

2023). Understanding these applications is essential, as they form the foundation upon which XAI techniques are applied to enhance interpretability.

2.2 Explainable AI (XAI) Techniques in Medical Imaging

This category encompasses studies that focus on the development and application of XAI methods to elucidate the decision-making processes of AI models in medical imaging. These papers provide insights into various techniques for making AI models more interpretable and trustworthy in medical contexts. Grad-CAM introduced a technique that provides visual explanations for deep networks by highlighting important regions in input images. It utilizes gradients flowing into the final convolutional layer to produce localization maps, aiding in understanding model decisions (Selvaraju et al., 2017). An extension of Grad-CAM, Grad-CAM++ improved localization of object regions in images, especially when multiple instances of the same class are present. It refined the weighting of pixels, resulting in more precise and interpretable heatmaps (Chattopadhay et al., 2018). Score-CAM proposed a gradient-free approach to generate visual explanations by weighting activation maps with the model's output scores. This method enhanced the robustness and interpretability of explanations without relying on gradient information (Wang et al., 2020).

LIME was introduced to explain the predictions of any classifier by approximating it locally with an interpretable model. It worked by perturbing the input data and observing the changes in the output, thereby identifying the contribution of each feature to the prediction (Ribeiro et al., 2016). SHAP unified several previous methods to assign each feature an importance value for a particular prediction. Based on cooperative game theory, SHAP computed Shapley values that fairly distributed the "payout" (model prediction) among the features. In medical imaging, SHAP provided both global and local interpretability by quantifying the contribution of each pixel or region to the model's output. This facilitated a deeper understanding of model behavior. Despite its theoretical soundness, SHAP's computational complexity posed challenges for high-dimensional data like images (Lundberg & Lee, 2017). LRP was proposed to decompose the prediction of a neural network by propagating the prediction score backward through the network layers. This

method assigned relevance scores to each neuron, ultimately attributing the model's output to the input features. However, LRP's effectiveness depended on the network architecture and required careful tuning to produce meaningful explanations (Bach et al., 2015). Counterfactual explanations were introduced to provide insights into model decisions by identifying minimal changes to the input that would alter the prediction. This approach answered "what-if" scenarios, helping users understand the decision boundaries of complex models. But, generating realistic and plausible counterfactuals in high-dimensional image spaces remained a significant challenge (Wachter et al., 2017).

2.3 Research Objectives

The integration of deep learning models into medical imaging has significantly enhanced diagnostic capabilities. However, the opaque nature of these models often hinders their acceptance in clinical settings due to a lack of interpretability. To bridge this gap, XAI techniques have emerged as vital tools to elucidate the decision-making processes of complex models. This research aims to systematically investigate and implement XAI methods to enhance the transparency and trustworthiness of deep learning applications in medical image classification and segmentation. By doing so, the study seeks to facilitate the adoption of AI-driven tools in healthcare by making their outputs more understandable to medical professionals. The specific objectives of this research are:

- Systematically evaluate the effectiveness of various XAI techniques in enhancing the interpretability of deep learning models for medical image classification and segmentation.
- Develop guidelines for integrating XAI methods into clinical workflows to facilitate the adoption of AI models in medical practice.
- Assess the impact of the XAI techniques on clinician trust and decision-making.

By achieving these objectives, we aim to make AI tools in healthcare more transparent and trustworthy, ultimately improving patient care.

CHAPTER 3: RESEARCH METHODS

The main goal of this research is to explore how XAI techniques can enhance the interpretability and trustworthiness of deep learning models used in medical image classification and segmentation. To achieve this, we will evaluate various XAI methods and apply them to benchmark datasets in the context of medical imaging. By providing more transparent and understandable explanations for AI decisions, we aim to bridge the gap between AI models and healthcare professionals, ensuring that AI tools can be confidently used in clinical practice.

The research questions guiding this study are as follows:

- How do different XAI techniques compare in enhancing the interpretability of deep learning models for medical image classification and segmentation?
- What is the impact of these XAI techniques on clinicians' trust and decision-making processes?
- How can XAI methods be effectively integrated into clinical workflows to support diagnostic accuracy and efficiency?

3.1 Research Approach and Design

This research will employ a mixed-methods approach, combining both quantitative and qualitative analysis. The quantitative component will involve measuring the effectiveness of XAI methods using standard performance metrics, such as accuracy, precision, and recall, as well as metrics specifically designed to assess the quality of visual and feature-based explanations. The qualitative component will include feedback from medical professionals to assess the usefulness, reliability, and interpretability of the generated explanations.

The rationale for this mixed-methods approach is to provide a comprehensive evaluation of XAI techniques from both technical and practical perspectives. While the quantitative analysis will focus on the model performance and the accuracy of explanations, the qualitative analysis will

ensure that the results are aligned with the needs and expectations of clinicians who will ultimately use these AI tools.

3.2 Proposed Methodology

The proposed research will consist of the following key components:

- **Dataset Selection**: We will use publicly available medical image datasets, such as NIH ChestX-ray14 and BraTS for brain tumor segmentation, to evaluate the effectiveness of the XAI techniques. These datasets are well-suited for our research as they cover diverse imaging modalities and medical conditions.

- **Model Development**: We will train state-of-the-art deep learning models, such as U-Net for segmentation and ResNet for classification, on the selected datasets. These models have demonstrated strong performance in medical imaging tasks and will serve as the foundation for our XAI evaluations.

- **Application of XAI Methods**: We will apply various XAI techniques, including Grad-CAM, Grad-CAM++, Score-CAM, LRP, and SHAP, to the trained models. These methods will be used to generate visual and feature-based explanations that highlight the regions and features influencing the models' decisions.

- **Quantitative Evaluation**: We will use a set of predefined metrics, such as the fidelity of the explanations, localization accuracy, and computational efficiency, to assess the effectiveness of each XAI method in providing clear and reliable insights into the models' decision-making processes.

- **Qualitative Evaluation**: Medical experts will review the generated explanations to assess their interpretability and usefulness in clinical settings. We will gather feedback through surveys and interviews to understand how well the explanations help clinicians trust and use the AI models.

- **Integration into Clinical Workflows**: Based on the results of the quantitative and qualitative evaluations, we will explore ways to incorporate the selected XAI methods into existing clinical workflows. This will include recommendations for integrating visual explanations and feature attributions into radiology software or other medical imaging platforms.

The chosen research methods are designed to address both the technical and practical aspects of XAI in medical imaging. Deep learning models have shown excellent performance in medical image classification and segmentation, but their lack of transparency has been a major barrier to their adoption in clinical practice. By applying XAI techniques, we can make these models more interpretable, providing healthcare professionals with the necessary insights to trust AI-driven decisions. The combination of quantitative and qualitative methods will allow us to comprehensively assess both the technical effectiveness of the XAI methods and their usability in real-world medical settings.

CHAPTER 4: EXPECTED CONTRIBUTIONS

The primary goal of this research is to enhance the transparency and interpretability of deep learning models used in medical image classification and segmentation through XAI techniques. By applying and evaluating different XAI methods, this research aims to provide insights into how AI models make decisions, which will help improve clinician trust in AI-driven diagnostic tools. The findings from this research will not only contribute to advancing the field of medical AI but also provide practical solutions for integrating XAI methods into clinical workflows, thus making AI tools more accessible and reliable for healthcare professionals.

4.1 Scientific Merit

This research will make significant contributions to the scientific community in the following ways:

- **Advancement of XAI Techniques**: The evaluation and comparison of various XAI methods, including Grad-CAM, Grad-CAM++, Score-CAM, LRP, and SHAP, will deepen our understanding of their strengths and limitations in the context of medical image analysis. This study will provide a comprehensive assessment of how these methods perform across different medical imaging tasks, filling an important gap in the current literature on XAI applications in healthcare.

- **Improvement in AI Interpretability**: By applying these XAI techniques to state-of-the-art deep learning models, the research will help improve the interpretability of AI models used in medical image segmentation and classification. This will allow researchers and practitioners to better understand the factors influencing AI predictions, promoting more transparent and reliable AI models in healthcare.

- **Benchmarking XAI in Medical Imaging**: The research will establish benchmarks for the use of XAI in medical imaging, providing guidelines and best practices for researchers looking to implement these methods. This will be valuable for future studies and for those seek-

ing to develop more explainable AI solutions for medical applications.

Through these contributions, this research will advance both the field of XAI and its practical applications in medical imaging, pushing the boundaries of what is known about the intersection of interpretability and AI-driven diagnostics.

4.2 Broader Impacts

This research has the potential to contribute significantly to society, particularly in the fields of healthcare and AI ethics:

- **Improvement in Healthcare Decision-Making**: By making AI models more interpretable and trustworthy, this research will help healthcare professionals better understand AI-driven diagnostic tools. This increased understanding will lead to more confident decision-making, ultimately improving patient outcomes, especially in critical areas like early disease detection and personalized treatment planning.

- **Bridging the Gap Between AI and Clinical Practice**: One of the key contributions of this research is the development of practical solutions for integrating XAI methods into clinical workflows. By ensuring that AI models are transparent and interpretable, this work will facilitate the broader adoption of AI in healthcare, helping to reduce the skepticism that many clinicians have regarding AI technologies.

- **Ethical AI and Fairness in Healthcare**: The transparency provided by XAI techniques will help ensure that AI models are not only effective but also fair and accountable. This is especially important in healthcare, where AI systems must be transparent to avoid potential biases in decision-making. This research aligns with the growing emphasis on ethical AI and the need for fairness, accountability, and transparency in AI systems used in high-stakes applications like healthcare.

Overall, the broader societal impact of this research lies in its potential to make AI in healthcare more transparent, ethical, and trustworthy, leading to improved patient care, greater acceptance

of AI tools by healthcare providers, and a more equitable and accountable AI-driven healthcare system.

REFERENCES

Arrieta, A. B., DÃaz-RodrÃguez, N., Ser, J. D., Bennetot, A., Tabik, S., Barbado, A., ... Herrera, F. (2020). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, *58*, 82-115. Retrieved from https://www.sciencedirect.com/science/article/pii/S1566253519308103 doi: https://doi.org/10.1016/j.inffus.2019.12.012

Bach, S., Binder, A., Montavon, G., Klauschen, F., MÃŒller, K.-R., & Samek, W. (2015, 07). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, *10*(7), 1-46. Retrieved from https://doi.org/10.1371/journal.pone.0130140 doi: 10.1371/journal.pone.0130140

Chattopadhay, A., Sarkar, A., Howlader, P., & Balasubramanian, V. N. (2018). Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 ieee winter conference on applications of computer vision (wacv)* (p. 839-847). doi: 10.1109/WACV.2018.00097

Chen, H., Gomez, C., Huang, C.-M., & Unberath, M. (2022). Explainable medical imaging ai needs human-centered design: guidelines and evidence from a systematic review. *NPJ digital medicine*, *5*(1), 156.

Farrag, A., Gad, G., Fadlullah, Z. M., Fouda, M. M., & Alsabaan, M. (2023). An explainable ai system for medical image segmentation with preserved local resolution: Mammogram tumor segmentation. *IEEE Access*, *11*, 125543-125561. doi: 10.1109/ACCESS.2023.3330465

Holzinger, A. (2018). From machine learning to explainable ai. In *2018 world symposium on digital intelligence for systems and machines (disa)* (p. 55-66). doi: 10.1109/DISA.2018.8490530

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems* (p. 4768â4777). Red Hook, NY, USA: Curran Associates Inc.

Mahmood, T., Rehman, A., Saba, T., Nadeem, L., & Bahaj, S. A. O. (2023). Recent advancements and future prospects in active deep learning for medical image segmentation and classification. *IEEE Access*, *11*, 113623-113652. doi: 10.1109/ACCESS.2023.3313977

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (p. 1135â1144). New York, NY, USA: Association for Computing Machinery. Retrieved from `https://doi.org/10.1145/2939672.2939778` doi: 10.1145/2939672.2939778

Roby, M., Restrepo, J. C., Park, H., Muluk, S. C., Eskandari, M. K., Baek, S., & Finol, E. A. (2025). Automatic segmentation of abdominal aortic aneurysm from computed tomography angiography using a patch-based dilated unet model. *IEEE Access*, *13*, 24544-24554. doi: 10.1109/ACCESS.2025.3533417

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In N. Navab, J. Hornegger, W. M. Wells, & A. F. Frangi (Eds.), *Medical image computing and computer-assisted intervention – miccai 2015* (pp. 234–241). Cham: Springer International Publishing.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 ieee international conference on computer vision (iccv)* (p. 618-626). doi: 10.1109/ICCV.2017.74

Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, *31*, 841.

Wang, H., Wang, Z., Du, M., Yang, F., Zhang, Z., Ding, S., ... Hu, X. (2020). Score-cam: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition workshops* (pp. 24–25).

Xie, Y., Zhang, J., Lu, H., Shen, C., & Xia, Y. (2021). Sesv: Accurate medical image segmentation by predicting and correcting errors. *IEEE Transactions on Medical Imaging*, *40*(1), 286-296. doi: 10.1109/TMI.2020.3025308

Zhou, T., Fu, H., Zhang, Y., Zhang, C., Lu, X., Shen, J., & Shao, L. (2020). M2net: Multi-modal multi-channel network for overall survival time prediction of brain tumor patients. In *International conference on medical image computing and computer-assisted intervention.* Retrieved from `https://api.semanticscholar.org/CorpusID:219791943`