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 Layer-wise Relevance Propagation (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, consis-

tency, 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 Explainable AI (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. Understanding these applications is essential, as they form the foundation upon which XAI techniques are applied to enhance interpretability. The related papers are: (Farrag et al., 2023; Mahmood et al., 2023; Roby et al., 2025; Ronneberger et al., 2015; Xie, Zhang, Lu, Shen, & Xia, 2021; Zhou et al., 2020)

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 are directly related to the research as they provide insights into various techniques for making AI models more interpretable and trustworthy in medical contexts. The related papers are: (Bach et al., 2015; Chattopadhay et al., 2018; Chen et al., 2022; Holzinger, 2018; Lundberg & Lee, 2017; Ribeiro et al., 2016; Selvaraju et al., 2017; Wachter et al., 2017; Wang et al., 2020)

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`