"A Comparative Study on Explainable AI Models for Medical Image Diagnosis"

Mrs.S.Kasthuri,

Teaching Assistant,Surya Engineering College,Erode.

Abstract: In this research, we dive into a comparative study of Explainable Artificial Intelligence (XAI) models used in medical image diagnosis. While AI models, particularly deep learning, have shown impressive results in spotting abnormalities in medical images, their lack of transparency can be a real concern when it comes to clinical decision-making. This study takes a closer look at various XAI techniques—like LIME, SHAP, Grad-CAM, and Integrated Gradients—across different deep learning models, utilizing medical imaging datasets such as ChestX-ray14 and HAM10000. We assess metrics like accuracy, interpretability, and how usable these models are for clinicians. The findings shed light on the balance between model performance and explainability, providing valuable insights for choosing the best XAI approach in clinical environments.

Keywords: Medical Image Diagnosis, Deep Learning, Model Interpretability, Grad-CAM, SHAP

INTRODUCTION

Over the past few years, how artificial intelligence (AI) has been integrated into the healthcare industry has simply revolutionized healthcare, particularly the diagnosis of medical images. With increasing access to large, labeled medical image datasets and advances in computing power, deep learning algorithms—specifically Convolutional Neural Networks (CNNs)—have demonstrated impressive ability in detecting and classifying radiological image patterns such as X-rays, MRIs, CT scans, and histopathological images. These developments have the potential to greatly improve diagnostic accuracy, reduce human error, and improve clinical efficiency. But even though these AI models performed so well, their "black-box" character poses a very serious problem in critical fields such as healthcare, where understanding and transparency are indispensable.

Doctors tend to struggle to put their trust in AI systems that cannot provide transparent, understandable reasons for their actions. A false diagnosis can be fatal. It's for this reason that Explainable Artificial Intelligence (XAI) is the growing area of interest—a set of methods meant to make AI systems more transparent and explainable decision-making processes clearer for humans. XAI is essential in bridging human clinicians with AI systems by offering graphic or textual interpretations which can support and legitimize automated choices. It is particularly crucial in the field of medicine, where professional, legal, and ethical guidelines mandate that support for clinical decisions be understandable and traceable.

Although numerous studies have separately used these methods in medical imaging tasks, a systematic comparison of their strengths and weaknesses, and hence their real-world usefulness, is still not available. In this work, an attempt is made to fill this void by comparing various XAI models for explaining deep learning-based diagnostic systems. We target extensively employed datasets like ChestX-ray14 (chest disease detection) and HAM10000 (skin lesion classification) that are typical of actual clinical image datasets. The datasets contain diverse pathological conditions and serve as standard benchmarks in medical AI work.

The research employs deep learning frameworks like CNNs and blended models like ResNet and InceptionV3 as base classifiers. The diagnostic models' performance is measured by conventional metrics like accuracy, precision, recall, and F1-score, whereas the models' explainability is measured in terms of visual clarity, clinical relevance, and usability as rated by domain specialists. This double evaluation system enables us to not only realize which models are best for the task of prediction but also which ones are most helpful and reliable from the clinician's viewpoint. This study adds to the emerging body of evidence in ethical and human-centered AI in healthcare. Systematically comparing and assessing XAI methods in a medical context, the work offers insightful recommendations for researchers, developers, and healthcare practitioners who want to deploy AI in clinical processes. It underlines that medicine's AI needs to transcend accuracy and adopt transparency, accountability, and cooperation with human decision-makers.

RESEARCH QUESTION

1. Which XAI model produces the most trustworthy visual explanations for different imaging modalities like X-rays, CT scans, or dermoscopic images?
2. How do clinicians perceive and analyze the explanations given by different XAI methods in a clinical setting?
3. What are the compromises between model performance (accuracy, precision, recall) and interpretability in using XAI in medical image diagnosis?

LITERATURE REVIEW:

The growing use of artificial intelligence (AI) in medicine, particularly medical image analysis, has created a corresponding need for transparent explainable AI (XAI) models that provide decision-making transparency. A survey of recent publications shows a mounting concern regarding the “black-box” characteristics of deep learning algorithms and emphasizes the clinical importance of interpretability.

Convolutional Neural Networks (CNNs) have been found effective for tasks like tumor detection, skin lesion classification, and lung disease diagnosis. Litjens et al. (2017) presented an extensive review that summarized more than 300 deep learning tasks in medical image analysis, proving the efficacy of CNNs in abnormality detection. Even though these models perform excellently, they fail to offer intuitive explanations, restricting their acceptance among medical professionals.

Samek et al. (2019) and Holzinger et al. (2017) contend that interpretability in AI is fundamental for trust and regulation in medicine. In the absence of transparency, AI systems are prone to misdiagnosis and malpractice. Therefore, incorporating XAI frameworks is no longer a choice but an imperative for health systems. Different XAI models have been proposed to explain intricate deep learning outputs. Ribeiro et al. (2016) presented LIME, which is a model explanation technique that approximates local predictions of a model with simpler models. Lundberg and Lee (2017) proposed SHAP, which applies Shapley values to determine feature importances. In image tasks, Grad-CAM (Selvaraju et al., 2017) and Integrated Gradients (Sundararajan et al., 2017) have been popularized through creating visual heatmaps that identify the regions of an image that affect the prediction. They provide varying levels of interpretability, with Grad-CAM tending to be the favorite in medical imaging because of the simplicity of its visual feedback.

Efforts have been made to assess the utility of XAI methods. Arun et al. (2021) compared Grad-CAM and Integrated Gradients on chest X-rays and concluded that Grad-CAM provided more clinically useful heatmaps. The research was restricted to a single dataset and lacked feedback from end-users, clinicians. Tjoa and Guan (2020) surveyed XAI applications in medical imaging and highlighted the absence of common benchmarks and comparative frameworks.

THEORETICAL FRAMEWORK

The rapid growth of Artificial Intelligence (AI) in the healthcare sector, particularly in medical imaging, has led to the development of impressive diagnostic tools that can sometimes surpass human experts in specific tasks. However, these AI systems, especially deep learning models like Convolutional Neural Networks (CNNs), often operate as black boxes, leaving us in the dark about how they arrive at their decisions. This lack of clarity has sparked skepticism and hesitation in the clinical world, underscoring the pressing need for Explainable AI (XAI). The theoretical foundation of this study pulls from various fields, including machine learning, medical informatics, cognitive trust, and human-computer interaction, to evaluate how different explainable AI models perform in diagnosing medical images.

At the heart of this framework is deep learning—particularly CNNs—which have shown remarkable success in image-related tasks like classification, segmentation, and detection in medical imaging (think X-rays, MRIs, and CT scans). The core principle behind CNNs involves hierarchical feature extraction, where different layers identify increasingly complex patterns. Despite their impressive performance, CNNs lack interpretability, which complicates their use in clinical settings where transparency is crucial for medical accountability.

EXPLAINABLE ARTIFICIAL INTELLIGENCE (XAI)

XAI offers both theoretical and practical tools to help us understand the decisions made by AI models. Several theoretical frameworks support XAI, including: Feature Attribution Models: Techniques like SHAP (SHapley Additive exPlanations), LIME (Local Interpretable Model-agnostic Explanations), and Integrated Gradients provide scores or visual representations that illustrate how each feature (like pixels in an image) influences a decision.

INTERPRETATION OF EXPLAINABILITY METHODS

The use of Explainable AI (XAI) in medical image diagnosis marks a crucial advancement in building trust between artificial intelligence and clinical practice. This study explored various XAI techniques applied to deep learning models for medical imaging, emphasizing interpretability, accuracy, and clinical relevance. The results offer valuable insights that can enhance both academic research and real-world applications in healthcare AI systems. The study revealed a common trade-off between model accuracy and interpretability. More complex models, such as deeper CNN architectures or ensemble models, often achieved higher diagnostic accuracy but sacrificed some explainability. This highlights the importance of finding a balance between predictive performance and transparency—especially in sensitive fields like healthcare, where explainability is vital for decision-making, legal accountability, and ensuring patient safety.

From a clinical perspective, the subjective assessment of explanation clarity and usefulness indicated that visual explanation techniques were generally favored over text-based or feature-importance charts. This preference aligns with the inherently visual nature of radiological diagnostics. However, clinicians also

stressed the importance of having consistent and reliable explanations across different cases, as inconsistent outputs could undermine trust in AI systems.

CONCLUSION

In recent years, we've seen a remarkable shift in how Artificial Intelligence (AI) is being woven into medical diagnostics, leading to notable improvements in both the efficiency and accuracy of disease detection. This is especially true with the use of deep learning models in analyzing medical images. However, the "black box" nature of these AI systems still poses a significant hurdle for their broader acceptance in clinical settings. This research tackled that issue by conducting a comparative study of various Explainable AI (XAI) techniques, with the goal of bridging the gap between high-performing AI models and the essential need for transparency, interpretability, and trust in healthcare.

The study took a close look at several cutting-edge XAI methods—including Grad-CAM, LIME, SHAP, and Integrated Gradients—using popular medical imaging datasets. These techniques were assessed based on their ability to deliver meaningful, clear, and clinically relevant explanations for the predictions made by deep learning models. It became clear that while each method has its own strengths and weaknesses, visual explanation techniques like Grad-CAM stood out for image-based diagnoses, as they effectively highlighted the areas of interest directly on the images. Conversely, methods like SHAP provided strong interpretability for structured data, helping to shed light on the overall behavior of the models.

One of the key takeaways from this research is that there isn't a one-size-fits-all XAI method that excels in every diagnostic task. Instead, the best choice of explanation model really depends on the specific medical context, the type of image being analyzed, and the diagnostic objectives at hand. This underscores the importance of a task-driven approach when selecting the right XAI technique. Additionally, the findings pointed out a trade-off between model complexity and interpretability: while more complex models often deliver greater accuracy, they can be harder to interpret, whereas simpler models, while easier to explain, might not perform as well.

REFERENCE

1. Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(4), e1312. https://doi.org/10.1002/widm.1312
2. 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* (pp. 4765–4774). Curran Associates Inc.
3. 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* (pp. 1135–1144). https://doi.org/10.1145/2939672.2939778
4. 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 *Proceedings of the IEEE International Conference on Computer Vision* (pp. 618–626). https://doi.org/10.1109/ICCV.2017.74
5. Tjoa, E., & Guan, C. (2020). A survey on explainable artificial intelligence (XAI): Toward medical XAI. *IEEE Transactions on Neural Networks and Learning Systems*, 32(11), 4793–4813. https://doi.org/10.1109/TNNLS.2020.3027314

6. Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., 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. https://doi.org/10.1016/j.inffus.2019.12.012
7. Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., ... & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42, 60–88. https://doi.org/10.1016/j.media.2017.07.005