International Journal of Engineering Research and Development
e- ISSN: 2278-067X, p-ISSN: 2278-800X, www.ijerd.com
Volume 22, Issue 6 (June 2026), PP 33-36

Deep Learning-Based Medical Image Analysis Using Explainable Artificial Intelligence

Nikita Joshi
Department of Computer Science
Institute of Management Studies Ghaziabad(University Courses Campus)

ABSTRACT: *The rapid advancement of deep learning techniques has increased the demand for explainable models, particularly in critical domains such as medical image analysis where decision transparency is essential. This survey provides a comprehensive review of explainable artificial intelligence (XAI) approaches applied to deep learning-based medical image analysis. A structured XAI framework is proposed to categorize these methods based on specific explainability criteria. Existing research on XAI techniques in medical imaging is examined and organized according to the proposed framework as well as different anatomical regions. Finally, the survey highlights future research directions and emerging opportunities for XAI in medical image analysis.*

Date of Submission: 27-05-2026 Date of acceptance: 05-06-2026

I. INTRODUCTION

Deep learning has invoked tremendous progress in automated image analysis. Before that, image analysis was commonly performed using systems fully designed by human domain experts. For example, such image analysis system could consist of a statistical classifier that used handcrafted properties of an image (i.e., features) to perform a certain task. Features included low-level image properties such as edges or corners, but also higher-level image properties such as the spiculated border of a cancer. In deep learning, these features are learned by a neural network (in contrast to being handcrafted) to optimally give a result (or output) given an input. An example of a deep learning system could be the output 'cancer' given the input of an image showing a cancer. Neural networks typically consist of many layers connected via many nonlinear intertwined relations. Even if one is to inspect all these layers and describe their relations, it is unfeasibly to fully comprehend how the neural network came to its decision. Therefore, deep learning is often considered a 'black box'. Concern is mounting in various fields of application that these black boxes may be biased in some way, and that such bias goes unnoticed. Especially in medical applications, this can have far-reaching consequences. There has been a call for approaches to better understand the black box. Such approaches are commonly referred to as intepretable deep learning or explainable artificial intelligence (XAI) (Adadi and Berrada, 2018; Murdoch et al., 2019). These terms are commonly interchanged; we will use the term XAI. Some notable XAI initiatives include those from the United States Defense Advanced Research Projects Agency (DARPA), and the conferences on Fairness, Accountability, and Transparency by the Association for Computing Machinery (ACM FAccT). The stakes of medical decision making are often high. Not surprisingly, medical experts have voiced their concern about the black box nature of deep learning (Jia et al., 2020), which is the current state of the art in medical image analysis (Litjens et al., 2017; Meijering, 2020; Shen et al., 2017). Furthermore, regulations such as the European Union's General Data Protection Regulation (GDPR, Article 15) require the right of patients to receive meaningful information about how a decision was rendered. Researchers in medical imaging are increasingly using XAI to explain the results of their algorithms. Something can be considered a good explanation if it gives insight into how a neural network came to its decision and/or can make the decision understandable. In this survey, we aim to give a comprehensive overview of papers using XAI in medical image analysis. We chose to focus solely on papers that used deep learning-based XAI in medical image. Since XAI techniques often originate from computer vision, we will elaborate on papers that adapted XAI techniques from computer vision by adding domain knowledge from the medical imaging field. The papers are grouped in the tables according to explanation method and according to anatomical location. This survey adds to the review of Reyes et al. (2020); since they mainly discussed techniques in computer vision, without extensively evaluating the adaptation of such techniques throughout medical image analysis. Furthermore, we describe if and how techniques from computer vision have been adapted specifically for medical image analysis. This survey adds to the review of Huff et al. (2021), since they mostly focused on examples of visual explanation, while our survey aims for a more holistic approach including non-visual explanation, critiques on XAI, and methods for evaluating XAI.

II. EXPLAINABLE ARTIFICIAL INTELLIGENCE (XAI) FRAMEWORK

In this section, we will give a brief overview of Explainable Artificial Intelligence (XAI) techniques found in deep learning for medical image analysis. For exhaustive surveys focused solely on XAI, please refer to Adadi and Berrada (2018) and Murdoch et al. (2019). We will distinguish XAI techniques based on three criteria: model-based versus post hoc, model-specific versus modelagnostic, and global versus. The framework of these three criteria is adapted from the surveys of Adadi and Berrada (2018) and Murdoch et al. (2019)

A fundamental distinction in explainable artificial intelligence (XAI) is between model-based explanations and post hoc explanations.

2.1.1 Model-Based Explanation

Model-based explanations rely on inherently interpretable models, such as linear regression and support vector machines. These models are designed to be simple enough for human understanding while still effectively capturing relationships between input and output variables (Murdoch et al., 2019). Such methods are typically associated with conventional machine learning approaches.

Interpretability in model-based methods is generally achieved either by limiting the number of features involved (sparsity) or by ensuring that the model's entire decision-making process can be easily understood by humans (simulatability) (Murdoch et al., 2019). For example, the Least Absolute Shrinkage and Selection Operator (LASSO) technique (Tibshirani, 1996) reduces many coefficients to zero, retaining only a small set of significant features that influence the prediction. This enhances transparency and makes the model's functioning easier to interpret.

2.1.2 Post Hoc Explanation

Post hoc explanations are applied after a model, such as a deep neural network, has already been trained in order to understand the patterns and relationships it has learned. In contrast to model-based approaches, post hoc techniques do not require interpretability to be integrated into the model during training. Instead, they focus on explaining the behavior of complex "black-box" systems after development.

The major distinction is that post hoc methods aim to interpret the predictions generated by an existing neural network, whereas model-based approaches incorporate interpretability directly into the structure and design of the model itself.

2.2 Model-Specific and Model-Agnostic Explanations

Another important classification in XAI is the division between model-specific and model-agnostic explanation techniques. Although these concepts are closely related to model-based and post hoc explanations (Adadi and Berrada, 2018), they differ in certain aspects.

2.2.1 Model-Specific Explanation

Model-specific explanation methods are tailored for particular types of machine learning models or neural network architectures. These approaches take advantage of characteristics unique to a specific model category.

One drawback of model-specific methods is that they limit the choice of models to only those compatible with the selected explanation technique. Consequently, other potentially more effective or accurate neural network architectures may not be considered because they cannot be explained using that approach.

2.2.2 Model-Agnostic Explanation

Model-agnostic explanation techniques are independent of any specific neural network architecture. These methods analyze only the relationship between the model's inputs and outputs. By altering or perturbing input data and examining the corresponding changes in output, they identify the features or regions that most strongly affect the model's predictions.

Since model-agnostic methods evaluate trained models externally rather than embedding interpretability within the model design, they are commonly categorized as post hoc explanation approaches.

III. XAI IN MEDICAL IMAGE ANALYSIS

In this section, we will present which XAI techniques are used in medical image analysis, and we will discuss adaptations of the methods typically seen in computer vision. We categorize the explanation methods into three types: visual, textual, and example-based; and we will classify each method according to the framework of model-based versus post hoc, model-specific versus model-agnostic, and global versus local explanation (Fig. 1). Table 1 provides an overview of the most frequently used techniques and shows their connections according to the taxonomy defined in Section 2.

3.1. Visual explanation

Visual explanation, also called saliency mapping, is the most common form of XAI in medical image analysis (Fig. 2). Saliency maps show the important parts of an image for a decision. Most saliency mapping techniques use backpropagation-based approaches, but some use perturbation-based or multiple instance learning-based approaches. These approaches will be discussed below. An overview of papers using saliency maps in medical imaging is shown in Table 2

3.2. Textual explanation T

Textual explanation is a form of XAI that provides textual descriptions. Such descriptions include relatively simple characteristics (e.g. 'spiculated mass'), up to entire medical reports. We will describe three types of textual explanation: image captioning, image captioning with visual explanation, and testing with concept attribution.

3.3. Example-based explanation

Example-based explanation is an XAI technique that provides examples relating to the data point that is currently being analyzed. This can be useful when trying to explain why a neural network came to a decision, and is related to how humans reason. For example, when a pathologist examines a biopsy of a patient that shows similarity with an earlier patient examined by the pathologist, the clinical decision may be enhanced by knowing the assessment of that earlier biopsy. Example-based explanation often optimizes the hidden layers deep in the neural network (i.e., the latent space) in such a way that similar points are close to each other in this latent space, while dissimilar points are further away in the latent space.

IV. PROS AND CONS OF XAI TECHNIQUES

All XAI techniques described in Section 3 have pros and cons, influencing how one would choose from the various options. We will structure these pros and cons in the categories ease of use, validity, robustness, computational cost, necessity to fine-tune, and open-source availability. An overview of these pros and cons per method from Table 1 is given in Table 5.

4.1. Ease of use

We define the ease of use by the potential of XAI techniques to be 'plug-and-play'. Post hoc model agnostic techniques have the highest ease of use. These methods generally consist of perturbation-based visual explanation techniques such as occlusion sensitivity. These techniques can be used on any trained neural network to provide a visual explanation. Model-based techniques typically have lowest ease of use, since the explanation is embedded in the design of the neural network.

4.2. Validity

We define validity by whether the explanation is correct and corresponds to what the end-user expects. In case of visual explanation, this can be assessed for example by asking a radiologist whether the explanation points towards the pathology that the neural network was designed to classify. Research on quantifying validity of XAI is sparse, and currently focuses on visual explanation. Arun et al. (2021) aimed to quantify the validity of visual explanation techniques using the SIIMACR Pneumothorax Segmentation and RSNA Pneumonia Detection databases (Society for Imaging Informatics in Medicine and American College of Radiology, 2019; Radiological Society of North America, 2018). They compared four of the methods discussed in this paper: backpropagation, guided backpropagation, Grad-CAM, and guided Grad-CAM. Of these methods, Grad-CAM showed the highest validity. Note that this study solely focuses on chest X-rays. Therefore, more research is needed to investigate the validity of visual explanation techniques in other modalities and anatomical locations.

4.3. Robustness

The robustness of XAI techniques can be assessed by intentionally changing certain aspects of the deep learning framework and measuring the effect of these changes to the given explanation. The robustness is mainly quantified for visual explanation techniques, using parameter randomization tests and data randomization tests. The parameter randomization test compares visual explanation from a trained CNN with visual explanation from a randomly initialized untrained CNN of the same architecture. If the explanation depends on the learned parameters of the CNN (the desired situation), the two explanations should differ substantially. If the two explanations are similar, the visual explanation technique is insensitive to the properties of the CNN

4.4. Computational cost

Computational cost of XAI is seldom reported in papers, but can be assessed by comparing how these explanation techniques work. Since model-based techniques embed the explanation in the design of the neural network, it is obvious that these explanations are relatively costly to produce. For visual explanation techniques, there is a clear distinction between backpropagation-based and perturbationbased techniques with respect to their

computational needs. Backpropagation-based techniques typically make a single pass back through the neural network, which is relatively fast. Perturbation-based techniques require, however, extensive perturbation of input images to measure the influence of these perturbations on the output. Therefore, these techniques are generally more computationally-expensive. This can especially be the case in 3-dimensional, 4-dimensional, and/or multi-modality images, which often occur in medical image analysis. The computational costs of the post hoc textual explanation TCAV and the post hoc example-based explanation of influence functions in medical image analysis has not rigorously been reported.

4.5. Necessity of fine-tuning

Some explanation techniques require no fine-tuning of parameters while others require fine-tuning of parameters associated with the XAI technique. Since model-based techniques embed the explanation in the design of the neural network, it is obvious that fine-tuning of the network will influence the explanation. For visual explanation, most backpropagation techniques have a limited number of parameters to tune. For example, in Grad-CAM, the user needs to choose at which layer to inspect the activation and in Deep SHAP, one needs to choose samples from the training set to calculate a background signal.

V. DISCUSSION

We derived our XAI framework from the frameworks of Adadi and Berrada (2018) and Murdoch et al. (2019). Other frameworks also exist, such as the framework by Kim et al. that divides XAI in pre-, during-, and post-model explanation. Duringand post-model explanation are captured by our XAI framework with model-based and post hoc explanation. Pre-model explanation mainly focuses on the structure of a dataset, such as inspecting outliers. One could state that an example-based explanation that utilizes the latent distributions of a dataset could be perceived as a pre-model explanation. We have, however, not made this distinction, since in deep learning, these latent distributions are discovered by training a neural network. We tried to be as comprehensive as possible with the inclusion of papers in our survey. However, XAI often is a technique used to support methods, and keywords are often not mentioned in the title or body of papers (Rudin, 2019). Therefore, we cannot guarantee that we covered all the work in the field. Nevertheless, we provided the search strategy to be as transparent as possible about the selection of papers.

VI. CONCLUSION

This paper surveyed research papers using explainable artificial intelligence (XAI) in deep-learning based medical image analysis, classified according to an XAI framework, and categorized according to anatomical location and imaging technique. The paper discussed how to evaluate XAI, current critiques on XAI, and future perspectives for XAI in medical image analysis.

REFERENCES

[1] Arun, N., et al. (2021). Assessing the validity of saliency maps for explainable AI in medical imaging. Medical Image Analysis.

[2] Adadi, A., Berrada, M., 2018. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access 6, 52138–52160. doi:10.1109/ACCESS. 2018.2870052.

[3] D. T. Huff, A. J. Weisman, R. Jeraj, 2021. Interpretation and visualization techniques for deep learning models in medical imaging. *Physics in Medicine & Biology*, 66, 04TR01.

[4] Jia, X., et al. (2020). Explainable AI in medical imaging: Current status and future directions. Frontiers in Artificial Intelligence.

[5] Murdoch, W.J., Singh, C., Kumbier, K., Abbasi-Asl, R., Yu, B., 2019. Definitions, methods, and applications in interpretable machine learning. Proc. Natl. Acad. Sci. USA 116, 22071–22080. doi:10.1073/pnas.1900654116

[6] Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J.A.W.M., van Ginneken, B., Sánchez, C.I., 2017. A survey on deep learning in medical image analysis. Med. Image Anal. doi:10.1016/j.media.2017.07.005.

[7] M. Rezaei, T. Uemura, J. Näppi, H. Yoshida, C. Lippert, C. Meinel, 2020. Generative synthetic adversarial network for internal bias correction and handling class imbalance problem in medical image diagnosis. *Medical Imaging 2020: Computer-Aided Diagnosis*, 113140E.

[8] Meijering, E., 2020. A bird's-eye view of deep learning in bioimage analysis. Comput. Struct. Biotechnol. J. doi:10.1016/j.csbj.2020.08.003.

[9] Shen, D., Wu, G., Suk, H.I., 2017. Deep learning in medical image analysis. Annu. Rev. Biomed. Eng. 19, 221–248. doi:10.1146/annurev-bioeng-071516-044442.

[10] Van de Schoot, R., de Bruin, J., Schram, R. *et al.* An open source machine learning framework for efficient and transparent systematic reviews. *Nat Mach Intell* **3**, 125–133 (2021). https://doi.org/10.1038/s42256-020-00287-7