ISSN: 2454-9940

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

E-Mail :
editor.ijasem@gmail.com
editor@ijasem.org

www.ijasem.org

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT
ISSN 2454-9940
www.ijasem.org
Vol 14, Issue 3, 2020
https://zenodo.org/records/15835503

SURVEY OF EXPLAINABLE AI TECHNIQUES IN HEALTHCARE

[1]Shekhar Katukoori, [2]Dr.Sandeep Chahal

[1]Research Scholar, shekhar.ktk@gmail.com

Department of Computer Science and Engineering

[2]NIILM UNIVERSITY, Kaithal, Haryana, India. Associate Professor

Computer Science and Engineering, NIILM UNIVERSITY

Abstract

Artificial intelligence (AI) with deep learning models has been widely applied in numerous domains, including medical imaging and healthcare tasks. In the medical field, any judgment or decision is fraught with risk. A doctor will carefully judge whether a patient is sick before forming a reasonable explanation based on the patient's symptoms and/or an examination. Therefore, to be a viable and accepted tool, AI needs to mimic human judgment and interpretation skills. Specifically, explainable AI (XAI) aims to explain the information behind the black-box model of deep learning that reveals how the decisions are made. This paper provides a survey of the most recent XAI techniques used in healthcare and related medical imaging applications. We summarize and categorize the XAI types, and highlight the algorithms used to increase interpretability in medical imaging topics. In addition, we focus on the challenging XAI problems in medical applications and provide guidelines to develop better interpretations of deep learning models using XAI concepts in medical image and text analysis. Furthermore, this survey provides future directions to guide developers and researchers for future prospective investigations on clinical topics, particularly on applications with medical imaging.

Keywords: explainable AI; medical imaging; deep learning; radiomics

1. INTRODUCTION

Currently, artificial intelligence, which is widely applied in several domains, can perform well and quickly. This is the result of the continuous development and optimization of machine learning algorithms to solve many problems, including in the healthcare field, making the use of AI in medical imaging one of the most important scientific interests [1]. However, AI based on deep learning algorithms is not transparent, making clinicians uncertain about the signs of

diagnosis. The key question then is how one can provide convincing evidence of the responses. However, there exists a gap between AI models and human understanding, currently known as “black-box” [2] transparency. For this reason, many research works focus on simplifying the AI models for better understanding by clinicians, in order to improve confidence in the use of AI models [3]. For example, the Defense Advanced Research Projects Agency (DARPA) of the United States developed the explainable AI (XAI) model in 2015. Later, in 2021, a trust AI project showed that the XAI can be used in interdisciplinary types of application problems, including psychology, statistics, and computer science, and may provide explanations that increase the trust of users [4]. Typically, XAI is an explainable model providing insights into how the predictions are made to achieve trustworthiness, causality, transferability, confidence, fairness, accessibility, and interactivity [5,6]. For example, as shown in Figure 1, it is strongly recommended to allow the AI model to be understandable for the public when the model outputs a decision. It is noted that the definition of XAI is not clear enough according to [7]. In addition, the 1. Introduction Currently, artificial intelligence, which is widely applied in several domains, can perform well and quickly. This is the result of the continuous development and optimization of machine learning algorithms to solve many problems, including in the healthcare field, making the use of AI in medical imaging one of the most important scientific interests [1]. However, AI based on deep learning algorithms is not transparent, making clinicians uncertain about the signs of diagnosis. The key question then is how one can provide convincing evidence of the responses. However, there exists a gap between AI models and human understanding, currently known as “black-box” [2] transparency. For this reason, many research works focus on simplifying the AI models for better understanding by clinicians, in order to improve confidence in the use of AI models [3]. For example, the Defense Advanced Research Projects Agency (DARPA) of the United States developed the explainable AI (XAI) model in 2015. Later, in 2021, a trust AI project showed that the XAI can be used in interdisciplinary types of application problems, including psychology, statistics, and computer science, and may provide explanations that increase the trust of users [4]. Typically, XAI is an explainable model providing insights into how the predictions are made to achieve trustworthiness, causality, transferability, confidence, fairness, accessibility, and interactivity [5,6]. For example, as shown in Figure 1, it is strongly recommended to allow the AI model to be understandable for the public when the model outputs

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT
ISSN 2454-9940
www.ijasem.org
Vol 14, Issue 3, 2020
https://zenodo.org/records/15835503

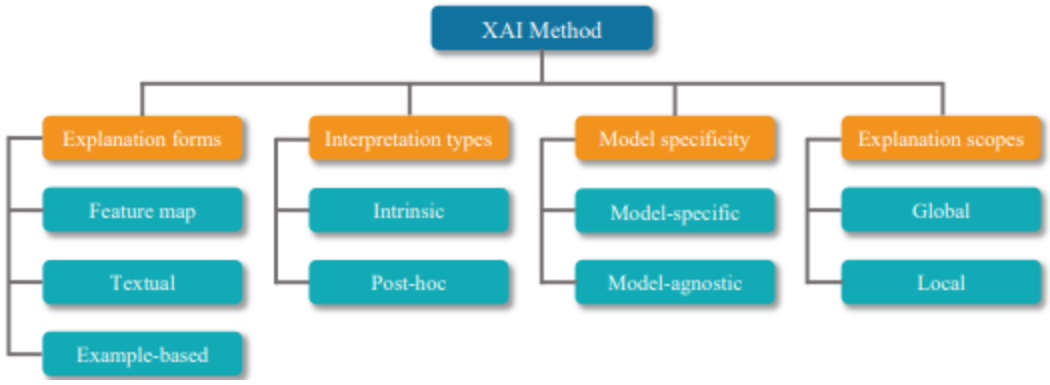
a decision. It is noted that the definition of XAI is not clear enough according to [7]. In addition, the

2. XAI Techniques Related to Medical Imaging

To trust AI models, the European Union has proposed seven key requirements, including (1) human agency and oversight; (2) technical robustness and safety; (3) privacy and data governance, (4) transparency; (5) diversity, non-discrimination, and fairness; (6) social and environmental well-being; and (7) accountability [12]. These seven requirements are summarized as follows.

3. Explainable Artificial Intelligence Techniques

This section provides a brief overview of the categories of XAI that can be used in healthcare. According to the literature published in recent years, there are many criteria used to classify XAI methods [25–27]. Figure 3 shows the criteria for classifying XAI methods and the corresponding categories. Based on these categories, it can be summarized that the most commonly used XAI techniques in medical fields are shown in Table 1. In addition, Table 2 reports the recent papers using the XAI method. For readability purposes, we have divided the table into explainable methods, modalities, and explanations of how explainable methods are applied. We have categorized the XAI techniques and explained in detail the methods as follows.



Categorization of explainable AI methods used in this paper. These criteria and categories are summarized from [25–27]. The orange and blue grids represent the criteria and categories, respectively

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT
ISSN 2454-9940
www.ijasem.org
Vol 14, Issue 3, 2020
https://zenodo.org/records/15835503

4. Introduction of the Explainable AI Method: A Brief Overview

As mentioned previously, XAI is widely used in many fields, in particular, medical imaging. In this section, we focus on the importance of XAI in healthcare applications.

4.1. Saliency Saliency directly uses the squared value of the gradient as the importance score of different input features [28]. The input can be graph nodes, edges, or node features. It assumes that the higher gradient value is related to the most important features. Although it is simple and efficient, it has several limitations. For example, it can only reflect the sensitivity between the input and output, which cannot express the importance very accurately. In addition, it has a saturation problem. For example, in regions where the performance model reaches saturation, the change in its output relative to any input change is very small, and the gradient can hardly reflect the degree of input contribution. Guided backpropagation (BP), whose principle is similar to that of the saliency map, modifies the process of backpropagating the gradient [29]. Since the negative gradients are hard to interpret, guided BP only back-propagates the positive gradients and shears the negative gradients to zero. Therefore, guided BP has the same limitations as saliency maps. One approach to avoid these limitations is to use layer-wise relevance propagation (LRP) [31] and deep Taylor decomposition (DTD) [74]. LRP and DTD are capable of improving a model's interpretability. In DTD, neural networks use complex non-linear functions that are represented by a series of simple functions. In LRP, the relevance of each neuron in the network is propagated backward through the network, thereby allowing it to quantify the contribution of each neuron to the final output. There are several rules designed with a specific type of layer in a neural network [31,74]. To combine LRP and DTD, LRP can be thought of as providing the framework for propagating relevance through a network, whereas DTD provides the means for approximating the complex non-linear functions used by the network. LRP and DTD may lead to overcoming the limitations of saliency maps and provide more accurate explanations [75].

4.1. Saliency Saliency directly uses the squared value of the gradient as the importance score of different input features [28]. The input can be graph nodes, edges, or node features. It assumes that the higher gradient value is related to the most important features. Although it is simple and efficient, it has several limitations. For example, it can only reflect the sensitivity between the

input and output, which cannot express the importance very accurately. In addition, it has a saturation problem. For example, in regions where the performance model reaches saturation, the change in its output relative to any input change is very small, and the gradient can hardly reflect the degree of input contribution. Guided backpropagation (BP), whose principle is similar to that of the saliency map, modifies the process of backpropagating the gradient [29]. Since the negative gradients are hard to interpret, guided BP only back-propagates the positive gradients and shears the negative gradients to zero. Therefore, guided BP has the same limitations as saliency maps. One approach to avoid these limitations is to use layer-wise relevance propagation (LRP) [31] and deep Taylor decomposition (DTD) [74]. LRP and DTD are capable of improving a model's interpretability. In DTD, neural networks use complex non-linear functions that are represented by a series of simple functions. In LRP, the relevance of each neuron in the network is propagated backward through the network, thereby allowing it to quantify the contribution of each neuron to the final output. There are several rules designed with a specific type of layer in a neural network [31,74]. To combine LRP and DTD, LRP can be thought of as providing the framework for propagating relevance through a network, whereas DTD provides the means for approximating the complex non-linear functions used by the network. LRP and DTD may lead to overcoming the limitations of saliency maps and provide more accurate explanations [75]. this is one of the most commonly used algorithms, it has some challenges. For example, the network's structure requires more flexibility so that the fully connected layer may adapt to the global average pooling layer. For this reason, a new algorithm known as "GradientCAM" is proposed to optimize CAM. It uses gradients to compute weight values [33]. First, the network is propagated forward to obtain the feature layer A (e.g., output of the last convolutional layer) and the network predicted value Y (e.g., output value before softmax activation).

4.3. Occlusion Sensitivity When training a neural network for image classification, the aim is to know whether this model can locate the position of the main target in the image. By partially occluding the picture, one can observe the situations of the network in the middle layers and the change in the predicted value after inputting the modified image. This leads to an understanding of why the network makes certain decisions. So far, occlusion sensitivity refers to how the probability of a given prediction changes with the occluded part(s) of the image. The higher the

output image value, the greater the decrease in the degree of certainty, indicating that the occlusion area is more important in the decision-making process [30]. 4.4. Testing with Concept Activation Vectors Testing with the concept activation vectors (TCAV) is an interpretable method proposed by the Google AI team [40]. Textual concepts are related to an explanation that is simple to understand. In the saliency map, it is not possible to explain the concept of pixels. For this reason, TCAV focuses on capturing high-level concepts in the neural network and attempts to provide a linear transformation from input to concepts using directional derivatives to quantify the importance of user-defined concepts to the classification results. However, this technique requires more investigation to be feasible in medical applications.

4.5. Triplet Networks The triplet network (TN) concept is an example-based framework [39]. For example, the TN training set consists of three samples: the first is randomly chosen from the "Anchor" training set, while the other two samples are randomly chosen from the training set in the same "Positive" and different "Negative" categories. By adjusting the parameters based on the distance between three inputs, the technique aims to bring the Anchor closer to the Positive and away from the Negative. Since labeling is not necessary, this method can be used for unsupervised learning. The technique is able to provide an explanation through the similarity between samples too.

REFERENCES

1. G. Valdes, J. M. Luna, E. Eaton, C. B. Simone, L. H. Ungar, and T. D. Solberg, "MediBoost: a patient stratification tool for interpretable decision making in the era of precision medicine," Scientific reports, vol. 6, no. 1, pp. 1-8, 2016.
2. H. Lakkaraju, E. Kamar, R. Caruana, and J. Leskovec, "Interpretable & explorable approximations of black box models," arXiv preprint arXiv:1707.01154, 2017.
3. A.B.Arrieta et al., "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," Information fusion, vol. 58, pp. 82-115, 2020.
4. A.B.Arrieta et al., "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," Information fusion, vol. 58, pp. 82-115, 2020.
5. A.B.Tosun, F. Pullara, M. J. Becich, D. Taylor, S. C. Chennubhotla, and J. L. Fine, "Histomapr™: An explainable ai (xai) platform for computational pathology solutions," in Artificial Intelligence and Machine Learning for Digital Pathology: Springer, 2020, pp. 204-227.

6. A.Das and P. Rad, "Opportunities and challenges in explainable artificial intelligence (xai): A survey," arXiv preprint arXiv:2006.11371, 2020.

7. A.Gomolin, E. Netchiporouk, R. Gniadecki, and I. V. Litvinov, "Artificial intelligence applications in dermatology: where do we stand?," Frontiers in medicine, vol. 7, p. 100, 2020.

8. A.Gupta, A. Anpalagan, L. Guan, and A. S. Khwaja, "Deep learning for object detection and scene perception in self-driving cars: Survey, challenges, and open issues," Array, vol. 10, p. 100057, 2021.

9. A.Gupta, A. Anpalagan, L. Guan, and A. S. Khwaja, "Deep learning for object detection and scene perception in self-driving cars: Survey, challenges, and open issues," Array, vol. 10, p. 100057, 2021.

10. A.Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller, "Causability and explainability of artificial intelligence in medicine," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 9, no. 4, p. e1312, 2019.

11. A.Khamparia et al., "Diagnosis of Breast Cancer Based on Modern Mammography using Hybrid Transfer Learning," Multidimensional Systems and Signal Processing, 2020.

12. A.Khamparia et al., "Diagnosis of Breast Cancer Based on Modern Mammography using Hybrid Transfer Learning," Multidimensional Systems and Signal Processing, 2020.

13. A.Polino, R. Pascanu, and D. Alistarh, "Model compression via distillation and quantization," arXiv preprint arXiv:1802.05668, 2018.

14. A.Rajkomar et al., "Scalable and accurate deep learning with electronic health records," NPJ Digital Medicine, vol. 1, no. 1, pp. 1-10, 2018.

15. A.Rajkomar et al., "Scalable and accurate deep learning with electronic health records," NPJ Digital Medicine, vol. 1, no. 1, pp. 1-10, 2018.

16. A.Rajkomar, J. Dean, and I. Kohane, "Machine learning in medicine," New England Journal of Medicine, vol. 380, no. 14, pp. 1347-1358, 2019.

17. Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE access, 6, 52138-52160.

18. Alicioglu, G., & Sun, B. (2022). A survey of visual analytics for Explainable Artificial Intelligence methods. Computers & Graphics, 102, 502-520.

19. Angelov, P. P., Soares, E. A., Jiang, R., Arnold, N. I., & Atkinson, P. M. (2021). Explainable artificial intelligence: an analytical review. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 11(5), e1424.

20. Ayano, Y. M., Schwenker, F., Dufera, B. D., & Debelee, T. G. (2022). Interpretable machine learning techniques in ECG-based heart disease classification: a systematic review. Diagnostics, 13(1), 111.

21. B.Letham, C. Rudin, T. H. McCormick, and D. Madigan, "Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model," The Annals of Applied Statistics, vol. 9, no. 3, pp. 1350-1371, 2015.

22. B.Mahbooba, M. Timilsina, R. Sahal, and M. Serrano, "Explainable artificial intelligence (xai) to enhance trust management in intrusion detection systems using decision tree model," Complexity, vol. 2021, 2021.

23. B.Mahbooba, M. Timilsina, R. Sahal, and M. Serrano, "Explainable artificial intelligence (xai) to enhance trust management in intrusion detection systems using decision tree model," Complexity, vol. 2021, 2021.

24. B.N.Patro, M. Lunayach, S. Patel, and V. P. Namboodiri, "U-cam: Visual explanation using uncertainty based class activation maps," presented at the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, 2019.

25. Banegas-Luna, A. J., Peña-García, J., Iftene, A., Guadagni, F., Ferroni, P., Scarpato, N., ...& Pérez-Sánchez, H. (2021). Towards the interpretability of machine learning predictions for medical applications targeting personalised therapies: A cancer case survey. International Journal of Molecular Sciences, 22(9), 4394.

26. C.Molnar, G. Casalicchio, and B. Bischl, "Quantifying interpretability of arbitrary machine learning models through functional decomposition," 2019.

27. C.Roggeman, W. Fias, and T. Verguts, "Salience maps in parietal cortex: imaging and computational modeling," Neuroimage, vol. 52, no. 3, pp. 1005-1014, 2010.

28. C.Sudlow et al., "UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age," PLoS medicine, vol. 12, no. 3, p. e1001779, 2015.

29. C.Xiao, T. Ma, A. B. Dieng, D. M. Blei, and F. Wang, "Readmission prediction via deep contextual embedding of clinical concepts," PloS one, vol. 13, no. 4, p. e0195024, 2018.