International Journal of Multiphysics
Volume 18, No. 4, 2024
ISSN: 1750-9548

Analysis of Explainable AI Methods in Healthcare

[1]Dr.Pathan Ahmed Khan, [2]Dr. Syed Amjed Hussaini
[1]Associate Professor, ISL Engineering College, drpathanahmedkhan@gmail.com
[2]Sr. Fire Protection Engineer, Reda Hazard Control , amjedhere@gmail.com

Abstract:
Artificial intelligence (AI) using deep learning models has been extensively used across several fields, including medical imaging and healthcare applications. In the medical domain, every judgment or choice is laden with danger. A physician will meticulously assess a patient's condition prior to developing a coherent explanation based on the patient's symptoms and/or an examination. Consequently, for AI to be a viable and acceptable instrument, it must replicate human judgment and interpretative abilities. Explainable AI (XAI) seeks to elucidate the underlying knowledge of deep learning's black-box models, clarifying the decision-making processes involved. This study presents an overview of the latest XAI approaches used in healthcare and associated medical imaging applications. We outline and classify the forms of XAI, emphasizing the techniques used to enhance interpretability in medical imaging subjects. Furthermore, we concentrate on the complex XAI issues within medical applications and provide guidance for enhancing the interpretability of deep learning models using XAI principles in medical picture and text analysis. This survey offers guidance for developers and researchers in future studies on clinical subjects, especially with medical imaging applications. A revolutionary transition towards Healthcare 5.0 is anticipated in the healthcare sector. It broadens the operational scope of Healthcare 4.0 and utilizes patient-focused digital wellbeing. Healthcare 5.0 emphasizes real-time patient monitoring, environmental management and wellbeing, as well as privacy adherence using assistive technologies such as artificial intelligence (AI), Internet of Things (IoT), big data, and supportive networking channels. Nonetheless, healthcare operational processes, the verifiability of predictive models, resilience, and the absence of ethical and legal frameworks pose possible obstacles to the actualization of Healthcare 5.0.

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

1. Introduction

Presently, artificial intelligence, extensively used across several fields, demonstrates high efficiency and rapid performance. This reflects the ongoing advancement and refinement of machine learning algorithms to address many challenges, particularly in healthcare, positioning the use of AI in medical imaging as a significant scientific focus [1]. Nonetheless, AI using deep learning algorithms lacks transparency, resulting in physicians' uncertainty over diagnostic indicators. The crucial inquiry is how to provide compelling proof of the replies. Nonetheless, a disparity persists between AI models and human comprehension, currently referred to as "black-box" transparency. Consequently, several research studies concentrate on streamlining AI models to enhance comprehension among doctors, thus increasing their trust in using these models [3]. In 2015, the Defense Advanced Research Projects Agency (DARPA) of the United States created the explainable AI (XAI) paradigm. Subsequently, in 2021, a trust AI initiative demonstrated that explainable artificial intelligence (XAI) may be used across several multidisciplinary domains, including psychology, statistics, and computer science, potentially offering explanations that enhance user trust [4].

XAI is an explicable model that elucidates the mechanisms behind predictions to foster trustworthiness, causality, transferability, confidence, fairness, accessibility, and interaction [5,6]. For instance, as seen in Figure 1, it is very advisable to ensure that the AI model's decision-making process is comprehensible to the public. The definition of XAI is deemed insufficiently explicit according to [7]. Furthermore, the terms “explainable” and “interpretable”

are linked to XAI concepts, whereby black-box models are deemed “explainable” when their predictions are analyzed using post hoc procedures. A model that is "interpretable" seeks to provide outputs that are comprehensible to humans in a stepwise manner [8]. The concept of explainability is contingent upon the prediction task, as shown in [9]. Consequently, the concept of explainability may be assessed according to the specific target users rather than by standardized criteria.

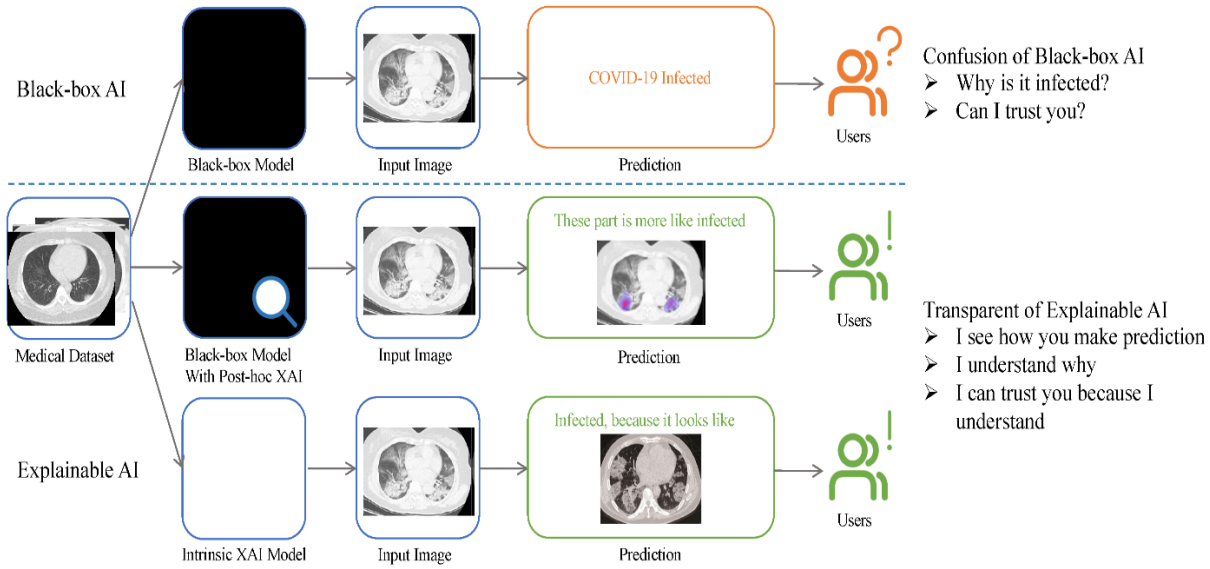

Figure 1. Flowchart of visual comparison between black-box and explainable artificial intelligence[1]

In healthcare 5.0, medical technology anticipates the integration of millions of IoT-based sensors that will transmit data over fifth-generation (5G) network infrastructure to enhance digital wellness, intelligent healthcare, and healthcare metrics. The integration of 5G, IoT, and AI creates a framework whereby intelligent mobile wearables are amalgamated with mobile communication and medical technologies to provide convenient and distant healthcare provision. Advanced IoT devices affixed to patients gather medical vitals, track progress, and diagnose health issues, transmitting data to physicians and medical institutions with little human intervention. 5G in IoT offers 10 Gbps speed, sub-10ms latency, secure future communications, expanded cellular coverage, improved network performance, and an approximate 90% improvement in battery longevity. AI algorithms, such as convolutional neural networks (CNN) and deep neural networks (DNN), execute intricate operations on extensive data sets, including image and text recognition, imaging, and provide precise illness prediction, detection, and remote healthcare treatment.

The main objective of the XAI model is to fostering confidence in the AI system among individuals. AI users may be categorized into two groups: (1) those possessing AI "expertise" and (2) individuals without such expertise. The first category pertains to specialists, algorithm creators, and scholars. They concentrate mostly on the AI model, devising novel techniques to oversee the information flow of an algorithm, as well as elucidating and enhancing its mechanisms. The second set of users mostly consists of domain specialists, including radiologists, and the general public. The specialist physicians need more elucidation about AI models to get a technological comprehension. Collaboration between academic and clinical researchers is advised.

Literature Review

The use of AI in healthcare has seen rapid progress in recent years, with research concentrating on its many applications and effects. Maleki and Forouzanfar [14] conducted a comprehensive examination of AI's potential in healthcare environments, highlighting its enhancements in diagnostic precision and patient care. Furthermore, recent research by Kalra et al. [19] investigated the expanding significance of AI in medical diagnosis, namely its incorporation into electronic health systems, and addressed the intricacies of integrating AI into clinical processes. Liu et al. [20] provided a comprehensive survey of AI applications in medicine, elucidating several AI models

and their efficacy in illness prediction and treatment.

Concurrently, XAI is an essential area in AI that focuses on the need for transparency and interpretability in AI models. Arrieta et al. [21] conducted a comprehensive evaluation of XAI strategies, classifying them into model-specific and model-agnostic categories, and emphasized the crucial significance of XAI in promoting trust and accountability in AI systems. Significant research in XAI encompasses Tosun et al. [22], which examined computational methodologies for XAI, presenting advanced tools for elucidation. Likewise, Longo et al. [23] presented a prospective view on XAI 2.0, delineating new interdisciplinary research avenues and tackling unresolved issues in AI model explainability.

Adadi and Berrada [24] examined the potential of XAI in several areas, including healthcare, highlighting fundamental obstacles in making AI models interpretable and suggesting viable solutions. The significance of Responsible AI, which includes ethical aspects like as justice, accountability, and transparency, has increased. Dignum [25] delineated concepts of Responsible AI, highlighting the need of integrating ethical norms with AI development methodologies. Mienye et al. [26] addressed fairness in AI, including methods for identifying and alleviating biases in healthcare machine learning algorithms.

Recent improvements have concentrated on model-specific XAI algorithms that provide explanations intricately embedded within the particular models used. An example is the research conducted by Konstantinov and Utkin [27], which presented novel techniques to enhance the interpretability of gradient-boosting machines via the use of parallel gradient boosting models. Their methodology employs linear combinations of boosting models and incorporates Lasso-based strategies to adjust model weights, making it very useful for diagnostic instruments, especially in fields like as cancer and cardiology. Furthermore, Raghavan [28] investigated the utilization of XAI in deep learning models tailored for medical imaging, wherever model-specific methodologies such as Grad-CAM and attention processes delivered instantaneous visual elucidations for MRI and CT scan evaluations.

A significant advancement in XAI is the emergence of hybrid XAI methods that integrate the advantages of both model-specific and model-agnostic methodologies. This amalgamation augments the adaptability and scalability of AI models across many healthcare sectors. Khan et al. [29] illustrated the use of hybrid XAI algorithms to structured and unstructured medical data, enhancing explainability in clinical decision support systems. These hybrid methodologies signify an emerging trend in XAI, overcoming the constraints of either model-agnostic or model-specific techniques by offering both global and local explanations, hence augmenting the transparency of AI systems used in intricate medical settings.

Holzinger et al. [30] emphasized the significance of human-in-the-loop (HITL) methodologies in artificial intelligence, especially within the healthcare sector. They contended that AI systems must be both accurate and capable of delivering explanations that are understandable to users in order to gain confidence and achieve widespread adoption in clinical practice. Their research highlights the importance of theoretical frameworks for XAI and the practical significance of HITL approaches in facilitating efficient AI integration.

Currently, the majority of available studies and surveys on AI and XAI provide a comprehensive overview of the domain, emphasizing particular applications or theoretical advancements. Singla [31] examined the use of AI in healthcare, offering insights on its potential but neglecting the interpretability issues encountered by healthcare workers. Esteva et al. [32] and Kaul et al. [33] similarly examined deep learning in healthcare, providing insights into its potential but neglecting the interpretability issues inherent to these models. This study seeks to provide a thorough overview of XAI in healthcare, including fundamental ideas, various applications, problems, and prospects for future research. This review seeks to address the existing vacuum in the literature and provide practical insights for academics, practitioners, and policymakers in the domain of healthcare AI.

Confidentiality and Privacy

AI systems need real-time data updates, presenting a systemic danger. The opaque nature of AI may lead to several security issues originating from both internal and external sources. These issues may pertain to the algorithm itself or external factors, such as user misuse and the generation of fraudulent datasets via network assaults [13,14]. In [15], the authors delineated three categories of security vulnerabilities associated with black-box AI: network assaults, system bias, and mismatch attacks. These dangers may result in significant repercussions for healthcare systems. Consequently, several research have examined and suggested solutions for XAI models regarding data security [16,17].

Ethics and Responsibilities

The medical field has established additional requirements for the application of AI, and clarifying the ethics and responsibilities associated with AI presents a significant challenge. Irresponsible AI may result in a reduction of medical personnel and patients [18]. Additionally, it encompasses the ethical considerations surrounding data privacy in the context of AI models. The current state of AI inspection and accountability is in its nascent phase. Further information regarding these issues and the interpretable functions of AI is available in [19,20]. The concept of responsible AI is analyzed to develop notions of responsibility within technological domains [21]. Explainability is a crucial factor in elucidating the responsibilities associated with AI.

2.3. Bias and Fairness

An AI model is trained on datasets with intrinsic qualities, which may include latent bias. In applications for age and skin color identification using photos of diverse ages and ethnicities, the AI model demonstrated a bias towards lighter skin tones and those aged 45 years [22]. Moreover, the inherent dangers in large data techniques that must not be overlooked facilitate heightened bias and repetition or amplify human mistakes [23]. The bias may originate from data, algorithms, and user interactions. This will significantly impact the equity of AI models, resulting in disparate outcomes for various individuals. In clinical medicine, various individuals may exhibit distinct symptoms that influence the algorithm. Consequently, it is essential to anticipate the effects of bias in AI algorithms within medical practice. This warrants thorough investigation when using AI in customized medicine.

3. Explainable Artificial Intelligence Techniques

A high-level summary of the several types of XAI with potential medical applications is given in this section. There are several criteria used to categorize XAI approaches, according to the research published in recent years. The classification criteria and associated categories for XAI approaches are shown in Figure 2. Table 1 displays the most popular XAI approaches utilized in medical domains, organized according to these categories. Table 2 also includes the most recent publications that have used the XAI approach. We have organized the table into sections for easier reading: explainable methodologies, modalities, and application descriptions.

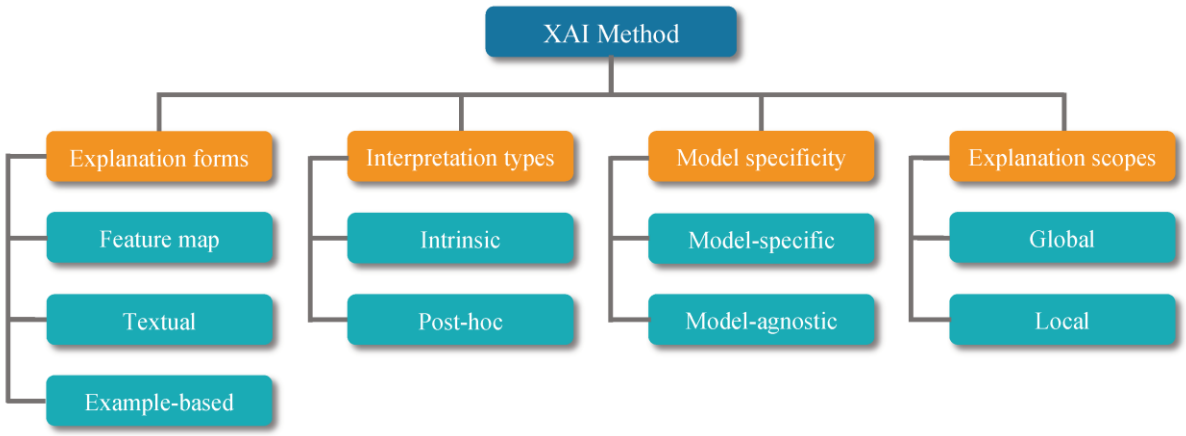

Figure 2. Categorization of explainable AI methods [1]

Introduction of the Explainable AI Method: A Brief Overview

A number of industries make use of XAI, but medical imaging is one of the most prevalent. The significance of XAI in healthcare applications is highlighted in this section.

4.1. Saliency

Saliency employs the squared gradient value as the significance score for various input characteristics [28]. The input may consist of graph nodes, edges, or node attributes. It posits that a greater gradient value correlates with the most significant aspects. Despite its simplicity and efficiency, it has several drawbacks. For instance, it might just indicate the sensitivity between the input and output, failing to convey the significance with precision.

Furthermore, it exhibits a saturation issue. In places where the performance model attains saturation, the output variation in response to any input alteration is minimal, rendering the gradient insufficient to accurately represent the extent of input contribution.

Guided backpropagation (BP), which operates on a premise similar to that of the saliency map, alters the gradient backpropagation process [29]. Due of the difficulty in interpreting negative gradients, guided backpropagation only propagates positive gradients while nullifying negative gradients. Consequently, guided backpropagation has the same constraints as saliency maps.

One method to circumvent these constraints is to use layer-wise relevance propagation (LRP) [31] and deep Taylor decomposition (DTD) [74]. LRP and DTD enhance a model's interpretability. In DTD, neural networks use intricate non-linear functions represented by a sequence of elementary functions. In LRP, the significance of each neuron in the network is sent backward, enabling the quantification of each neuron's contribution to the final output. Numerous regulations are formulated for a particular kind of layer inside a neural network [31,74]. LRP serves as the basis for disseminating relevance across a network, whereas DTD facilitates the approximation of the intricate non-linear functions used by the network. LRP and DTD may address the shortcomings of saliency maps and provide more precise explanations [75].Market Trends and Research Data

EXAI is becoming significant across many sectors, including healthcare, retail, marketing, media and entertainment, aerospace and defense, insurance, financial services, and the industrial Internet of Things (IIoT), among others. EXAI provides benefits such as enhanced client retention, improved inventory management, superior design interpretability, elevated performance and scalability, and decreased cost estimate. For instance, EXAI in the retail sector may forecast forthcoming fashion trends and enables businesses to showcase the newest items. In the e-commerce sector, EXAI may facilitate product searches based on its stored recommendations. In the formulation of corporate strategy, EXAI offers accountability and insights into essential business basics, including sales, consumer behavior patterns, and staff turnover, therefore reinforcing ethical business standards and mitigating bias and brand reputation damage. Recent market trends in EXAI indicate substantial benefits, including improved client retention, optimized management, and defect identification. In 2019, a third-party application developer compromised approximately 500 million Facebook accounts on Amazon Cloud Service during a fraud incident [23]. In 2020, a cyberattack targeted the core server of the National Highway Authorities of India (NHAI) because to inadequate cybersecurity architecture. In these situations, EXAI can elucidate the reasons behind such instances and provide measures to prevent future cyber assaults.

Methodology

The survey's design and methods adhere to the norms and restrictions established, The survey comprises six fundamental and analytical processes outlined in the following subsections.

A. Review Plan

The survey methodically emphasizes and delineates the issue description and outline. The essential aspects highlighted in the literature include (i) identification of the research question, (ii) analysis of data sources, research studies, and publications, (iii) logical parameters established for keyword searches relevant to the research, (iv) criteria for inclusion and exclusion, and (v) standardization and assessment of the search and research writing process.

B. Research Questions

The first phase of the survey involves articulating the research question to evaluate the survey's goals. This inquiry primarily addresses (i) the evolution and technological trends of EXAI in healthcare 5.0 applications, (ii) the seamless integration of technological advancements such as AI, 5G, and beyond (B5G) networks, along with FL in diverse applications to guarantee a quality user experience and interaction, and (iii) insights gained from the survey and the identification of future prospects within various human-centric applications.

C. Data Sources

The literature database employed for research is most relevant to computer science and medicine/healthcare. The literature explored IEEE Xplore, ACM Digital Library, PubMed etc., for research. This database provides a rich

source of information content. The study [35], [36] also recommends other electronic sources such as books, web blogs, preprints, articles, and patents for incorporation in the survey of interest.

Discussion, Challenges, and Prospects

Recent advancements in medical imaging mostly use post-interpretation techniques rather than model-based interpretation, with CAM and GCAM models being extensively utilized. These works concentrate on the implementation of algorithms, with interpretable techniques serving as an adjunct to the algorithms. In the lack of systematic advancements in XAI, there is a tendency to use local interpretation approaches to elucidate the examined situations. In the context of CNNs used to medical imaging, a saliency map serves as a straightforward instrument for elucidating the network's regions of interest [97]. Eight interpretable saliency map approaches, including Grad-CAM, guided backpropagation, and guided Grad-CAM, were assessed [19]. Nonetheless, the performance on the testing datasets was uncompetitive [19]. Consequently, several obstacles continue to confront the XAI approach.

Conclusion

This work has delineated some prominent XAI approaches about their concepts and implementation in medical imaging applications, including their efficacy. The algorithms were first categorized into many unique classifications. The use of AI in medical imaging has led to discussions over a prominent explainable AI (XAI) linked to medical picture classifications. A description of the applications of the newly suggested XAI methodologies to enhance the interpretability of their models was also provided. Moreover, the need for interpretable models in radiomic analysis was elucidated and examined. In conclusion, the discussion and analysis of the medical needs of XAI, together with its opportunities and obstacles for future exploration, were also presented. The survey delineates the fundamentals of EXAI, relevant metrics, and various implementations of EXAI use cases. The case study includes performance assessments that substantiate the advantages of EXAI in healthcare environments. The discussion encompasses the unresolved difficulties, research obstacles, and insights gained from the survey.

References:

1. Chaddad, A.; Peng, J.; Xu, J.; Bouridane, A. Survey of Explainable AI Techniques in Healthcare. Sensors 2023, 23, 634. https://doi.org/10.3390/s23020634
2. D. Saraswat et al., "Explainable AI for Healthcare 5.0: Opportunities and Challenges," in IEEE Access, vol. 10, pp. 84486-84517, 2022, doi: 10.1109/ACCESS.2022.3197671.
3. Nazar, M.; Alam, M.M.; Yafi, E.; Mazliham, M. A systematic review of human-computer interaction and explainable artificial intelligence in healthcare with artificial intelligence techniques. IEEE Access **2021**, 9, 153316–153348. [**Google Scholar**] [**CrossRef**]
4. von Eschenbach, W.J. Transparency and the black box problem: Why we do not trust AI. Philos. Technol. **2021**, 34, 1607–1622. [**Google Scholar**] [**CrossRef**]
5. Devlin, J.; Chang, M.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv **2018**, arXiv:1810.04805. [**Google Scholar**]
6. Gunning, D.; Aha, D. DARPA's explainable artificial intelligence (XAI) program. AI Mag. **2019**, 40, 44–58. [**Google Scholar**]
7. Angelov, P.; Soares, E. Towards explainable deep neural networks (xDNN). Neural Netw. **2020**, 130, 185–194. [**Google Scholar**] [**CrossRef**]
8. Arrieta, A.B.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; García, S.; Gil-López, S.; Molina, D.; Benjamins, R.; et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Inf. Fusion **2020**, 58, 82–115. [**Google Scholar**] [**CrossRef**] [**Green Version**]
9. Tjoa, E.; Guan, C. A survey on explainable artificial intelligence (xai): Toward medical xai. IEEE Trans. Neural Netw. Learn. Syst. **2020**, 32, 4793–4813. [**Google Scholar**] [**CrossRef**]

10. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat. Mach. Intell. **2019**, 1, 206–215. [**Google Scholar**] [**CrossRef**] [**Green Version**]
11. Gunning, D.; Stefik, M.; Choi, J.; Miller, T.; Stumpf, S.; Yang, G.Z. XAI—Explainable artificial intelligence. Sci. Robot. **2019**, 4, eaay7120. [**Google Scholar**] [**CrossRef**] [**Green Version**]
12. Yang, X.; He, X.; Zhao, J.; Zhang, Y.; Zhang, S.; Xie, P. COVID-CT-dataset: A CT scan dataset about COVID-19. arXiv **2020**, arXiv:2003.13865. [**Google Scholar**]
13. Falk, T.; Mai, D.; Bensch, R.; Çiçek, Ö.; Abdulkadir, A.; Marrakchi, Y.; Böhm, A.; Deubner, J.; Jäckel, Z.; Seiwald, K.; et al. U-Net: Deep learning for cell counting, detection, and morphometry. Nat. Methods **2019**, 16, 67–70. [**Google Scholar**] [**CrossRef**] [**PubMed**]
14. Maleki Varnosfaderani S., Forouzanfar M. The role of AI in hospitals and clinics: Transforming healthcare in the 21st century, Bioengineering, 2306-5354, 11 (4) (2024), 10.3390/bioengineering11040337
15. Smuha, N.A. The EU approach to ethics guidelines for trustworthy artificial intelligence. Comput. Law Rev. Int. **2019**, 20, 97–106. [**Google Scholar**] [**CrossRef**]
16. Bai, T.; Zhao, J.; Zhu, J.; Han, S.; Chen, J.; Li, B.; Kot, A. Ai-gan: Attack-inspired generation of adversarial examples. In Proceedings of the 2021 IEEE International Conference on Image Processing (ICIP), Anchorage, AK, USA, 19–22 September 2021; pp. 2543–2547. [**Google Scholar**]
17. Papernot, N.; McDaniel, P.; Jha, S.; Fredrikson, M.; Celik, Z.B.; Swami, A. The limitations of deep learning in adversarial settings. In Proceedings of the 2016 IEEE European symposium on security and privacy (EuroS&P), Saarbrücken, Germany, 21–24 March 2016; pp. 372–387. [**Google Scholar**]
18. Kiener, M. Artificial intelligence in medicine and the disclosure of risks. AI Soc. **2021**, 36, 705–713. [**Google Scholar**] [**CrossRef**] [**PubMed**]
19. Kalra N., Verma P., Verma S., Advancements in AI based healthcare techniques with FOCUS ON diagnostic techniques Comput Biol Med, 0010-4825, 179 (2024), Article 108917, 10.1016/j.compbiomed.2024.108917
20. Liu C., Tan Z., He M. Overview of artificial intelligence in medicine, Artificial intelligence in medicine, 9789811912221, Springer Nature Singapore (2022), pp. 23-34, 10.1007/978-981-19-1223-8_2
21. Barredo Arrieta A., Díaz-Rodríguez N., Del ,Ser J., Bennetot A., Tabik S., Barbado A., et al. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI Inf Fusion, 1566-2535, 58 (2020), pp. 82-115, 10.1016/j.inffus.2019.12.012
22. Vigano, L.; Magazzeni, D. Explainable security. In Proceedings of the 2020 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW), Genoa, Italy, 7–11 September 2020; pp. 293–300. [**Google Scholar**]
23. Kuppa, A.; Le-Khac, N.A. Black Box Attacks on Explainable Artificial Intelligence(XAI) methods in Cyber Security. In Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 19–24 July 2020; pp. 1–8. [**Google Scholar**] [**CrossRef**]
24. Trocin, C.; Mikalef, P.; Papamitsiou, Z.; Conboy, K. Responsible AI for digital health: A synthesis and a research agenda. Inf. Syst. Front. **2021**, 1–19. [**Google Scholar**] [**CrossRef**]
25. Verma, P. Bhattacharya, Y. Patel, K. Shah, S. Tanwar, and B. Khan, ‘‘Data localization and privacy-preserving healthcare for big data applications: Architecture and future directions,’’ in Emerging Technologies for Computing, Communication and Smart Cities, P. K. Singh, M. H. Kolekar, S. Tanwar, S. T. Wierzchoń, and R. K. Bhatnagar, Eds. Singapore: Springer, 2022, pp. 233–244.
26. T. Folke, S. C. Yang, S. Anderson, and P. Shafto, ‘‘Explainable AI for medical imaging: Explaining pneumothorax diagnoses with Bayesian teaching,’’ CoRR, vol. abs/2106.04684, pp. 1–20, Jun. 2021. [Online]. Available: https://arxiv.org/abs/2106.04684, doi: 10.48550/arxiv.2106.04684.
27. R.-X. Ding, I. Palomares, X. Wang, G.-R. Yang, B. Liu, Y. Dong, E. Herrera-Viedma, and F. Herrera, ‘‘Large-scale decision-making: Characterization, taxonomy, challenges and future directions from an artificial intelligence and applications perspective,’’ Inf. Fusion, vol. 59, pp. 84–102, Jul. 2020.
28. M. McFarland, ‘‘Uber shuts down self-driving operations in Arizona,’’ CNNMoney, vol. 809, p. 3, May 2018. [49] M. Bojarski, D. D. Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Müller, J. Zhang, X. Zhang, J. Zhao, and K. Zieba, ‘‘End to end learning for self-driving cars,’’ 2016, arXiv:1604.07316.

29. J. Haspiel, N. Du, J. Meyerson, L. P. Robert, Jr., D. Tilbury, X. J. Yang, and A. K. Pradhan, ‘‘Explanations and expectations: Trust building in automated vehicles,’’ in Proc. Companion ACM/IEEE Int. Conf. Hum.-Robot Interact. New York, NY, USA: Association for Computing Machinery, Mar. 2018, pp. 119–120.
30. X. Chao, ‘‘Behavior monitoring methods for trade-based money laundering integrating macro and micro prudential regulation: A case from China,’’ Technol. Econ. Develop. Econ., vol. 25, no. 6, pp. 1081–1096, 2019.
31. Sokol, K.; Flach, P. Explainability fact sheets: A framework for systematic assessment of explainable approaches. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, 27–30 January 2020; pp. 56–67.
32. Lipton, Z.C. The Mythos of Model Interpretability: In Machine Learning, the Concept of Interpretability is Both Important and Slippery. Queue 2018, 16, 31–57. [CrossRef]
33. Hoffman, R.R.; Mueller, S.T.; Klein, G.; Litman, J. Metrics for explainable AI: Challenges and prospects. arXiv 2018, arXiv:1812.04608.
34. Klein, G.; Hoffman, R.R. Macrocognition, mental models, and cognitive task analysis methodology. In Naturalistic Decision Making and Macrocognition; Ashgate Publishing: Farnham, UK, 2008; pp. 57–80.
35. Comiter, M. Attacking Artificial Intelligence AI’s Security Vulnerability and What Policymakers Can Do about It; Belfer Center for Science and International Affairs: Cambridge, MA, USA, 2019.
36. Druce, J.; Harradon, M.; Tittle, J. Explainable artificial intelligence (XAI) for increasing user trust in deep reinforcement learning driven autonomous systems. arXiv 2021, arXiv:2106.03775.
37. Le Merrer, E.; Trédan, G. Remote explainability faces the bouncer problem. Nat. Mach. Intell. 2020, 2, 529–539. [CrossRef]
38. Guang, Y.; Qinghao, Y.; Jun, X. Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond. Inf. Fusion 2022, 77, 29–52. [CrossRef]
39. Fauvel, K.; Masson, V.; Fromont, E. A performance-explainability framework to benchmark machine learning methods: Application to multivariate time series classifiers. arXiv 2020, arXiv:2005.14501.
40. Larochelle, H.; Erhan, D.; Courville, A.; Bergstra, J.; Bengio, Y. An empirical evaluation of deep architectures on problems with many factors of variation. In Proceedings of the International Conference on Machine Learning (ICML ‘07), Corvallis, OR, USA, 20–24 June 2007; pp. 473–480. [CrossRef]