International Journal of Sustainable Development Through AI, ML and IoT
ISSN(Online):2584-0827
Volume 4 | Issue 1 | 2025
https://ijsdai.com/index.php/IJSDAI/index

Explainable Artificial Intelligence (XAI) in healthcare: Addressing Techniques and Challenges

[1] Sumit [2] Dr Amrinder Kaur
[1] Research Scholar, Department of Computer Science & Application, Maharshi Dayanand University (Rohtak) sumit.rs24.dcsa@mdurohtak.ac.in
[2] Assistant Professor Department of Computer Science & Application, Maharshi Dayanand University (Rohtak) amrinder@mdurohtak.ac.in

ARTICLE INFO
Received: 01st June 2025
Revised: 17th June 2025
Accepted: 20th July 2025

ABSTRACT

The growing use of Artificial Intelligence (AI) models in high-stake applications like healthcare has driven the demand for transparency and explainability. This arises from the "black-box" nature of AI models because the wrong predictions from Artificial Intelligence (AI) can have high-impact consequences in crucial sectors like healthcare. The term Explainable Artificial Intelligence (XAI) includes the techniques and methodologies used to develop AI models that enable the users to comprehend the results and predictions generated by AI models. The success of XAI model integration within healthcare depends on its ability to be explainable and interpretable. Gaining the trust of healthcare professionals requires AI models to be more explainable and transparent regarding their outcomes. This paper provides an overview of XAI in healthcare, including techniques, challenges, opportunities, and emerging trends, to understand the realistic applications of XAI used in the field of healthcare. This study aims to discuss innovative perspectives and upcoming trends that can be useful to researchers and practitioners in adopting implementation of transparent and trustable AIdriven solutions in the healthcare sector.

Keywords: Explainable AI, Healthcare, Transparency, Interpretability, Explainability, Accountability

1. Introduction

In the constantly changing world of artificial intelligence (AI), Explainable Artificial Intelligence (XAI) stands out as the most trusted, legally compliant, effective, and powerful aspect of artificial intelligence. XAI encompasses frameworks and approaches capable of creating AI applications that can be understood by domain professionals or data scientists and by non-enthusiast of AI as well. The remarkable progress demonstrated by artificial intelligence (AI) in recent times, particularly its growing utilization in real-world applications, offers numerous opportunities for exploration and innovation. Typically, people tend to use AI models that are the most understandable. Such models able to integrate with AI systems are termed as transparent or explainable systems. Prior to digging in this field, there are two important concepts that are interrelated yet different. As one defines decisions, explainability considers how AI interactions happen within decisions, list the elements and present them in a way to figure out and break the explanation. Interpretability, on the other hand, deals with the human-understandable, understandable inflexible rules relative to a particular surrounding problem. There are multiple angles of viewing the value which lies in the explainability of AI.

Interpretability requires formulating a human-readable set of rules within a system's decision-making, while explainability is concerned with the construction of an interface through which a person can grasp how AI makes decisions. The value of AI explainability can be approached from so many angles. It entails designing a system interface that makes the workings of the AI's decision-making algorithm clear and understandable to humans (Arrieta et al., 2019). With advancements in the field of Artificial Intelligence (AI), their usage in the field of academic and industrial sectors have increased massively along with their side effects. In certain other fields like healthcare sectors, errors are not acceptable, where biological lives are at risk. The identification of the disease at an early stage with the highest degree of precision is essential for the patient's recovery or the prevention of the disease from progressing to more severe stages (Lesley and Hernandez, 2024).

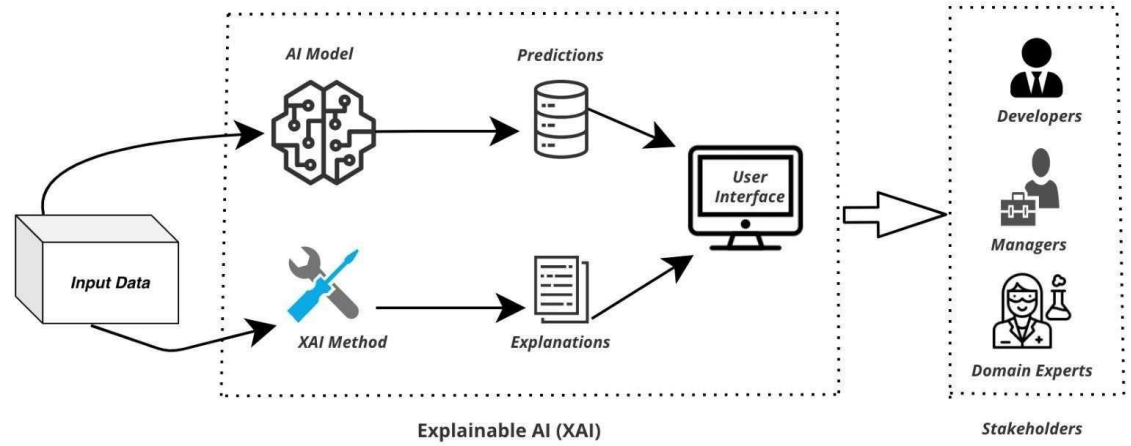

Figure 1: Explainable AI (XAI) Concept Design Architecture

2. Overview of XAI

XAI is an area of artificial intelligence that focuses on building models whose judgments humans can comprehend and interpret. Making the internal workings of AI systems visible and their results explicable is the main objective of XAI (Hassija et al., 2023). This is especially crucial in industries like healthcare, where knowing the logic behind AI choices may improve patient outcomes and increase confidence in the system.

The need of developing Explainable Artificial Intelligence models in such a way that the transparency of these models is easily understood by the human and enhance the trust in models' predictions. There are other additional goals of Explainable AI (XAI) as described:

1. Interpretability: The main objective of Explainable AI (XAI) is to make AI models and their outcomes clearly understandable to the users. This is essential to when AI is used to enhance the outcomes in the domains like healthcare.
2. Transparency: Transparency involves providing clear information about the internal operations of an AI model. This implies that people should be able to see and understand an AI model's internal operations, including the data it utilizes, the attributes it considers, and the reasoning it employs to arrive at a conclusion.
3. Understandability: The understanding describes the characteristics of a specific model that reflects its internal functioning to the users clearly – how that model is working and how its internal structure is processing data internally.
4. Comprehensibility: When the AI models are initiated with comprehensibility that describes an ability of a learning model to interpret its learned knowledge in a human understandable manner.
5. Trustworthiness: Building trust and confidence of users through intelligent and valid decisions taken by an AI model is the key component of trustworthiness. Trustworthiness can be attained by enhancing transparency, interpretability and stability of AI models that require the proper validation procedures and guidelines about models' decision-making process.

3. Taxonomy of XAI

Various techniques and methods are used for attaining the goals of XAI. These techniques are classified based on several factors that vary in the characteristics of the models.

- The XAI techniques are classified based on Interpretability: Perceptual Interpretability which are understandable by humans without any additional explanations and Interpretability through Mathematical frameworks which are based on the mathematical logics and equations (Tjoa and Guan, 2020).

- The further classification considers based on timing and integration, differentiating between ant-hoc models which provide explanations based on the internal mechanism of the model whereas on the other hand post-hoc models provide explanations when the model completes its learning (Holzinger et al., 2017).

- The other approaches are differentiated based on the targeted model: The approaches that do not depend on the internal architecture are Model-Agnostic approaches, on the other side model-specific approaches depend on the features of

a specific model (Combi et al., 2022).

- The other differentiation is between the fact of explanations is about the whole model or focused on single instances. Global Explanations which provide the explanations of the model as a whole whereas Local Explanations focus on the explanation of individual instances (Lundeberg et al., 2020).

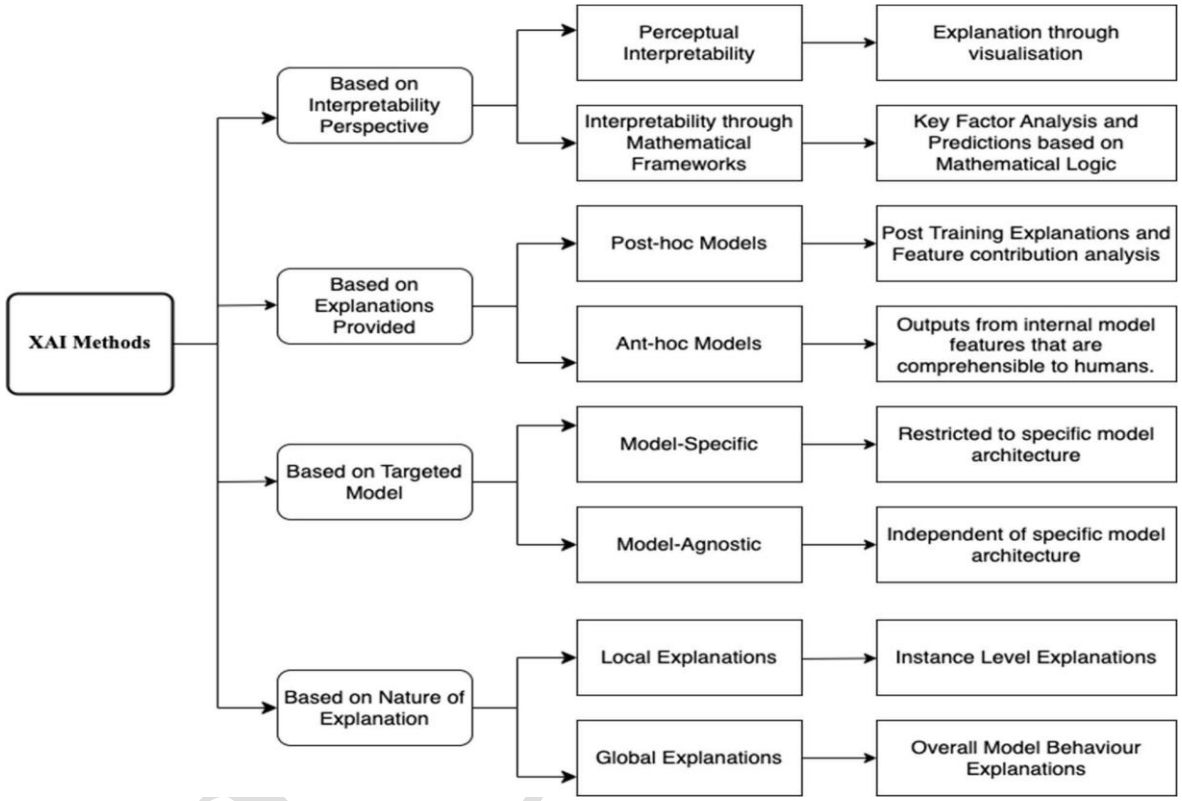

Fig 2: Flowchart of Taxonomies in Explainable Artificial Intelligence (XAI) approaches.

4. Techniques of XAI:

There are various techniques of Explainable Artificial Intelligence (XAI) that are used in healthcare applications depending upon the purpose of providing enhanced outcomes with accuracy and transparency. These techniques are mainly categorized into five important types based on their interpretation in AI models that are broadly used in healthcare application:

4.1. *Perturbation-based techniques:* These are the approaches which are used to examine AI models by altering the models' inputs such as text, images and other comparable data and navigate the changes in the outcomes of the model. It should be possible to determine which aspect of the input is most important for inference by observing the changes in the outcomes. The importance of the perturbated element is estimated by comparing the output when it is present or when it is absent (Ivanovs et al., 2021).

4.1.1. *Local Interpretable Model-Agnostic Explanations (LIME):* LIME is a popular perturbation technique of XAI that provides insights about the most important features influencing the output. It is a technique that provides a kind of local explanation which means it does not aim to explain the whole model by calculating all the inputs but focuses on explaining the individual instances. LIME only focused on the decision making process of any particular instance in that it wanted to use its classification based on the predictions. The instinct behind this technique is to make the "black-box" nature of an AI model to be explainable by approximating it through a simple interpretable model locally (Farhood et al., 2024).

4.1.2. *SHapley Additive Explanations (SHAP):* SHAP is an explainable AI technique used to describe feature importance concepts for specific predictions of any particular AI model. This technique is based on the cooperative game theory, where Shapley values are assigned to inputs to calculate the contribution of each input feature towards a model final predicted outcome. The benefit of using SHAP in healthcare is that it provides both local and global explanations, allowing users to understand the complete working of the model.

4.1.3. *The Randomized Input Sampling for Explanation (RISE):* The approach is to use the mathematical technique known as Monte Carlo sampling to build a random binary mask, followed by weighted averaging using the masked model's outputs. A saliency map is created by sampling many times and weighting the results. The result characterises the model's sensitivity to various areas of the input (Petsiuk et al., 2018).

4.2. *Backpropagation-based approaches:* Backpropagation-based approaches in Explainable AI use the gradients calculated during the training of the model to determine how the features of the input affect the output predictions. By reversing the output of the model through all the layers, these approaches determine which input feature is more vital for the decision-making process.

4.2.1. *Layer-wise Relevance Propagation approach (LRP*): As the name suggests, LRP is a backpropagation-based XAI technique used to explain the complex AI models, by reallocating the outcome scores from layer to layer with a backward pass through the network. LRP works in the backward direction by propagating each layer, where each input is assigned with a relevant score which describes its contribution towards the final outcomes of the model (Bach et al., 2015).

4.2.3. *Class Activation Mapping (CAM):* CAM essentially highlights the specific parts of an image that contribute to the final decision-making process of an AI model. Its structure closely resembles the Convolution Neural Networks. It consists of multiple convolution layers, with Global Average Pooling being performed by the layer immediately preceding the final output. By inputting the extracted features into the fully connected neural network layer controlled by the SoftMax activation function, the necessary probabilities are produced. This approach is useful in enhancing the explainability of the AI model, by visualising and clarifying the decision-making process (Yang et al., 2023).

4.3. *Gradient-based techniques:* These techniques examine how variations in input values affect the model's predictions, providing an understanding of feature significance and model dynamics. In gradient-based explainable AI, the connection between input modifications and shifts in output is measured to produce clear explanations of the model's predictions.

4.3.1. Saliency Map (SM): Saliency maps are considered a visualization method that

emphasizes the most important features of an input that affect the output of an AI model. This technique helps to illustrate which section of the input data plays a significant role in decision- making process of a particular AI model (Simonyan et al., 2013).

4.3.2. *Concept Activation Vectors (TCAV):* TCAV is a technique that uses the derivative directionals to measure the importance of the user-defined parameters to that of resulted outcomes. In contrast to other methods which are focused on input features, Concept Activation Vectors are focused on measuring the high-dimensional concepts like colours, race and gender to predict the importance of outcome class (Kim et al., 2025).

4.3.3. *Deep Learning Important Features (Deep-Lift):* Deep-Lift evaluates the importance of input features by comparing the activation levels of a neuron to those of a reference neuron. Deep-Lift employs the gradients data to calculate the influence or effect of individual input features on the output of the model, offering an in-depth insight into the internal activations and weights of the neural network. This allows users to clearly understand the reasoning behind decision-making of a particular models' process in specific intake situations.

4.3.4. *The Guided Backpropagation (GBP):* This method often referred to as guided saliency, is a variation of the deconvolution method for visualizing the features captured by CNNs and can be utilized across a wide array of network architectures. This method raises concerns about the application of max-pooling in convolutional neural networks for small images and suggests substituting max-pooling layers with a convolutional layer that has a greater stride, achieving no degradation in accuracy on various image recognition benchmarks.

4.4. *Instance-Based Techniques:* Instance-based explanations in Explainable AI (XAI) offer clarity on a model's decisions by highlighting particular instances from the dataset. Instead of delving into the model's internal workings, they assist users in grasping the rationale behind a choice by contrasting it with familiar and comparable choices.

4.4.1. *ANCHOR*: An anchor explanation refers to a rule that can successfully "anchor" a prediction to a particular local scenario relevant to the instance being examined. This implies that modifications made to other feature values of the instance won't considerably impact the rule's effectiveness in clarifying the prediction. The Anchors technique relies on reinforcement learning approaches and a graph search algorithm. It aims to reduce the number of model calls required during runtime while also efficiently recovering from local optima (AAAI, 2023).

4.4.2. *Multiple Instance Learning (MIL):* This technique can enhance the explainability of models. In MIL, data is grouped into bags, each consisting of several instances. While each bag is assigned a label, the individual instances within it lack specific labels. This arrangement allows for the identification of the instances that have the greatest impact on the label of the bag, thus improving model explainability. These are a type of weak supervised learning where the learning dataset contains a bag of instances (Fatima et al., 2023)

5. Applications of XAI in Healthcare:

XAI greatly improves multiple dimensions of healthcare by increasing the precision of

diagnoses, customizing treatments, providing transparent and interpretable insights understood by the doctors, patients and researchers. By ensuring the transparency and interpretability, XAI encourages the confidence and dependability in crucial sectors like healthcare that rely on AI.

5.1 Decision-Making Systems: The role of AI algorithms in Clinical Decision-Making are increasingly significant in the diagnosis and prediction of diseases, providing fresh perspectives in healthcare. These algorithms worked upon by examining the extensive volumes of patients' health data to uncover figures and relationships that are easily satisfiable to the humans.

5.2 *Medical Imaging:* Explainable Artificial Intelligence improves the comprehensibility of complex AI models used for highlighting images of internal body parts, which is vital for identifying and diagnosing conditions through X-rays, MRIs, and CT scans. Methods such as attention maps and Grad-CAM indicate the areas of the image that have the greatest impact on the model's decisions, offering radiologists visual insights that bolster their trust in AI results. Recent research has demonstrated the benefits of XAI across various applications in medical imaging (Salehi et al., 2023).

5.3 *Personalised Medicine:* XAI is essential in personalized medicine as it enhances the interpretability of machine learning models. These models evaluate genetic information, medical history, and lifestyle factors to recommend individualised treatment plans. XAI methods in healthcare provide understandable explanations for these suggestions, making sure therapies are customized to any individual's requirements, which in turn boosts both treatment outcomes and patient compliance (Ou et al., 2023).

5.4 *Remote Monitoring and Virtual Healthcare:* XAI improves the accuracy and attainability of remote monitoring of patients, particularly in the areas that are deprived or low accessible regions. By offering clear and transparent explanations with AI-powered diagnostic recommendations, XAI allows remote healthcare monitoring professionals to understand and have confidence in AI-powered recommendations, enhancing the quality of care in scenarios where access to healthcare specialists is limited (Albahri et al., 2023).

6. Obstacles and Prospects XAI in Healthcare:

Even with the considerable progress that has been achieved in creation of Explainable Artificial Intelligence (XAI), several ongoing issues continues to obstruct its effective and flawless deployment in healthcare sector. These challenges encompass technical, clinical, ethical as well as regulatory areas, and need to be resolved to guarantee that XAI tools and techniques are not only precise and understandable but also secure, reliable and practically applicable by the healthcare domain-experts in real-time scenarios.

6.1 *Concerns regarding data security:* The requirement of extensive amounts of sensitive patient information by the AI-powered systems in medical domains, it is crucial to guarantee the protection of this data (Carmody et al., 2021). The implementation of strong cybersecurity protocols and routine security assessments can enhance the security of data in safeguarding confidential information and building the trust between patients and healthcare professionals.

6.2 *Integration into existing healthcare infrastructures:* Incorporating XAI approaches into current healthcare infrastructural models can be rigorous results in domain-experts

(Doctors) may have to modify their steps and processes in integration of XAI models and techniques, which can be a lengthy and complicated undertaking. Ensuring the need of adopting the AI models into healthcare, there should be a user-friendly and alignment with the current systems.

6.3 *Data Sufficiency and Quality:* The quality and completeness of the data used for training the AI-powered systems plays a significant role in producing the qualitative outcomes. Data that is inconsistent or inaccurate can result in giving wrong outcomes from the systems. Therefore, the need of reliable and robust data of patients is a major challenge in the healthcare sector. So, it is necessary to gather data by using standardized methods and data curation processes to tackle this problem.

6.4 *Regulatory and compliance challenges:* Healthcare AI systems typically need authorization from regulatory agencies. This procedure can be time-consuming and complicated, as it necessitates thorough testing and validation of the AI models. Demonstrating that AI systems are safe and effective according to regulatory requirements poses a considerable challenge. It is essential to design the XAI models which follow these guidelines of ethical consideration rules without affecting the safety of patients, which is a crucial aspect of compliance.

7. Conclusion and Future Directions:

This research has found that a significant advantage of Explainable Artificial Intelligence (XAI) in the field of healthcare is the improvement of explaining the workings and decision making process of opaque AI models. Medical professionals need straightforward and comprehensible explanations for AI forecasts to guide their choices, particularly in high stakes situations like disease diagnosis or treatment plan formulation.

As a result, upcoming research avenues in XAI ought to concentrate on the following:

7.1 *Intelligent Robotic Systems for Surgical and Recovery Procedures:* The implementation of automation through the involvement of intelligent machines in surgical procedures is projected in producing significant strides (Feizi et al., 2023). Robotic systems, informed by AI technology, have the potential to carry out intricate surgeries with exceptional accuracy, thereby minimizing risks and enhancing patient results.

7.2 *Combining AI with IoT and Wearable Devices:* This fusion has the potential to create an analyzed record of data that is transmitted through providing health information of patients. This interconnectivity of AI and IoT also offers the necessary suggestions and notifications. Additionally, AI can be incorporated into IoT powered gadgets and technologies to enhance insights related to health and performance (Shi et al., 2020).

7.3 *Advancing models to increase understandability:* Models that are inherently interpretable should be a key focus for future research, aiming to enhance transparency from the ground up to make the complex AI models more understandable to the users resulting in strengthening the trust between patients' AI-powered diagnoses.

7.4 *Improving System Productivity:* The implementation of Explainable Artificial Intelligence in the healthcare sectors can overtake patient care and enhance system capacity and efficiency completely. In the future, innovations in AI-driven healthcare systems are expected to early disease detection, modify treatment plans and substantially personalize patient care, resulting in faster recovery rates and low mortality rates

References

AAAI. (2023, April 4). *Anchors: High-Precision Model-Agnostic Explanations - AAAI*. Retrieved April 25, 2025, from https://aaai.org/papers/11491-anchors-high-precision-model-agnostic-explanations

Albahri, A., Duhaim, A. M., Fadhel, M. A., Alnoor, A., Baqer, N. S., Alzubaidi, L., Albahri, O., Alamoodi, A., Bai, J., Salhi, A., Santamaría, J., Ouyang, C., Gupta, A., Gu, Y., & Deveci, M. (2023). A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion. *Information Fusion*, *96*, 156–191. https://doi.org/10.1016/j.inffus.2023.03.008

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2019). 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

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K., & Samek, W. (2015). On Pixel-Wise explanations for Non-Linear Classifier decisions by Layer-Wise relevance propagation. *PLoS ONE*, *10*(7), e0130140. https://doi.org/10.1371/journal.pone.0130140

Carmody, J., Shringarpure, S., & Van De Venter, G. (2021). AI and privacy concerns: A smart meter case study. *Journal of Information, Communication and Ethics in Society*, *19*(4), 492–505. https://doi.org/10.1108/JICES-04-2021-0042

Combi, C., Amico, B., Bellazzi, R., Holzinger, A., Moore, J. H., Zitnik, M., & Holmes, J. H. (2022). A manifesto on explainability for artificial intelligence in medicine. *Artificial Intelligence in Medicine*, *133*, 102423. https://doi.org/10.1016/j.artmed.2022.102423

Farhood, H., Najafi, M., & Saberi, M. (2024). Improving deep learning transparency: Leveraging the power of LIME Heatmap. In *Lecture notes in computer science* (pp. 72– 83). https://doi.org/10.1007/978-981-97-0989-2_7

Fatima, S., Ali, S., & Kim, H. (2023). A comprehensive review on multiple instance learning. *Electronics*, *12*(20), 4323. https://doi.org/10.3390/electronics12204323

Feizi, N., Tavakoli, M., Patel, R. V., & Atashzar, S. F. (2021). Robotics and AI for teleoperation, Tele-Assessment, and Tele-Training for Surgery in the era of COVID-19: existing challenges, and future vision. *Frontiers in Robotics and AI*, *8*. https://doi.org/10.3389/frobt.2021.610677

Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., Scardapane, S., Spinelli, I., Mahmud, M., & Hussain, A. (2023). Interpreting Black-Box Models: A review on Explainable Artificial intelligence. *Cognitive Computation*, *16*(1), 45–74. https://doi.org/10.1007/s12559-023-10179-8

Holzinger, A., Biemann, C., Pattichis, C. S., & Kell, D. B. (2017). What do we need to build explainable AI systems for the medical domain? *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1712.09923

Ivanovs, M., Kadikis, R., & Ozols, K. (2021). Perturbation-based methods for explaining deep neural networks: A survey. *Pattern Recognition Letters*, *150*, 228– 234. https://doi.org/10.1016/j.patrec.2021.06.030

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., & Sayres, R. (2017, November 30). *Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV)*. arXiv.org. Retrieved April 25, 2025, from https://arxiv.org/abs/1711.11279

Lesley, U., & Hernández, A. K. (2024). Improving XAI Explanations for Clinical Decision-Making – Physicians' perspective on local explanations in healthcare. In *Lecture notes in computer science* (pp. 296–312). https://doi.org/10.1007/978-3-031-66535-6_32

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, *2*(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9

Ou, S., Tsai, M., Lee, K., Tseng, W., Yang, C., Chen, T., Bin, P., Chen, T., Lin, Y., Sheu, W. H., Chu, Y., & Tarng, D. (2023). Prediction of the risk of developing end-stage renal diseases in newly diagnosed type 2 diabetes mellitus using artificial intelligence algorithms. *BioData Mining*, *16*(1). https://doi.org/10.1186/s13040-023-00324-2

Petsiuk, V., Das, A., & Saenko, K. (2018, June 19). *RISE: Randomized Input Sampling for explanation of black-box models*. arXiv.org. Retrieved April 25, 2025, from https://arxiv.org/abs/1806.07421

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": explaining the predictions of any classifier. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1602.04938

Salehi, A. W., Khan, S., Gupta, G., Alabduallah, B. I., Almjally, A., Alsolai, H., Siddiqui, T., & Mellit, A. (2023). A study of CNN and transfer learning in Medical imaging: Advantages, challenges, future scope. *Sustainability*, *15*(7), 5930. https://doi.org/10.3390/su15075930

Shi, Q., Dong, B., He, T., Sun, Z., Zhu, J., Zhang, Z., & Lee, C. (2020). Progress in wearable electronics/photonics—Moving toward the era of artificial intelligence and internet of things. *InfoMat*, *2*(6), 1131–1162.https://doi.org/10.1002/inf2.12122

Simonyan, K., Vedaldi, A., & Zisserman, A. (2013, December 20). *Deep inside convolutional networks: visualising image classification models and saliency maps*. arXiv.org. Retrieved April 25, 2025, from https://arxiv.org/abs/1312.6034

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

Yang, Z., Shao, J., & Yang, Y. (2023). An improved CycleGAN for data augmentation in person Re-Identification. *Big Data Research*, *34*, 100409. https://doi.org/10.1016/j.bdr.2023.10