A Survey of Explainable Artificial Intelligence in Healthcare: Concepts, Applications, and Challenges

Ibomoiye Domor Mienye , George Obaido * , Nobert Jere , Ebikella Mienye , Kehinde Aruleba , Ikiomoye Douglas Emmanuel , Blessing Ogbuokiri

Posted Date: 23 August 2024

doi: 10.20944/preprints202408.1702.v1

Keywords: AI; bias; ethics; fairness; healthcare; machine learning

Article

A Survey of Explainable Artificial Intelligence in Healthcare: Concepts, Applications, and Challenges

Ibomoiye Domor Mienye [1], George Obaido [2,*], Nobert Jere [3], Ebikella Mienye [1], Kehinde Aruleba [4], Ikiomoye Douglas Emmanuel [5] and Blessing Ogbuokiri [6]

[1] Institute for Intelligent Systems, University of Johannesburg, Auckland Park, Johannesburg 2006, South Africa; ibomoiyem@uj.ac.za (I.D.M.); 219105099@student.uj.ac.za (E.M.)
[2] Center for Human-Compatible Artificial Intelligence (CHAI), Berkeley Institute for Data Science (BIDS), University of California, Berkeley, CA 94720, USA
[3] Department of Computer Science, University of Fort Hare, Alice Campus, Alice, Eastern Cape, South Africa; njere@ufh.ac.za
[4] School of Computing and Mathematical Sciences, University of Leicester, Leicester LE1 7RH, UK; ka388@leicester.ac.uk
[5] School of Science, Engineering and Environment, University of Salford, Salford, UK; i.d.emmanuel@edu.salford.ac.uk
[6] Department of Computer Science, Brock University, Niagara Region, St. Catharines, ON L2S 3A1, Canada; bogbuokiri@brocku.ca
* Correspondence: gobaido@berkeley.edu

Abstract: Explainable AI (XAI) has the potential to transform healthcare by making AI-driven medical decisions more transparent, trustworthy, and ethically compliant. Despite its promise, the healthcare sector faces several challenges, including balancing interpretability and accuracy, integrating XAI into clinical workflows, and ensuring adherence to rigorous regulatory standards. This paper provides a comprehensive review of XAI in healthcare, covering techniques, challenges, opportunities, and advancements, thereby enhancing the understanding and practical application of XAI in healthcare. The study also explores responsible AI in healthcare, discussing new perspectives and emerging trends and providing valuable insights for researchers and practitioners. The insights and recommendations presented aim to guide future research and policy-making, promoting the development of transparent, trustworthy, and effective AI-driven solutions.

Keywords: AI; bias; ethics; fairness; healthcare; machine learning

1. Introduction

Artificial Intelligence (AI) has significantly reshaped numerous sectors, with healthcare witnessing significant transformations [1–3]. The integration of AI technologies into healthcare systems has led to advancements that were previously unattainable. These advancements include enhanced diagnostic accuracy, personalized treatment plans, and improved operational efficiencies [4–6]. Machine learning (ML) and deep learning (DL), as subfields of AI, enable the processing and analysis of vast datasets, which is crucial for predictive analytics and decision-making in clinical environments [7].

However, the adoption of AI in healthcare comes with challenges, particularly regarding the transparency and interpretability of AI models. Explainable AI (XAI) addresses these concerns by providing mechanisms that make AI decision-making processes comprehensible to humans. This is vital in healthcare, where understanding the rationale behind AI predictions is crucial to clinicians and patients [8–10]. Furthermore, as AI systems become more sophisticated, their decision-making processes become less transparent, raising concerns about trust, accountability, and ethical use. Addressing these issues through XAI can enhance the acceptance and reliability of AI technologies in medical practice.

In the past, researchers have conducted various reviews on AI in healthcare. Some reviews have centred on the general application of AI models in healthcare [9,11–13], while others have addressed

 doi:10.20944/preprints202408.1702.v1

the potential of AI for predictive analytics in clinical settings [14–16] and drug discovery [2,17,18]. Additionally, other reviews include discussions on the ethical implications and regulatory challenges associated with AI in healthcare [5,12,19]. However, most reviews either focus on AI broadly or do not provide an in-depth analysis of XAI tailored to healthcare. Additionally, over the years, there have been several advances in the field of healthcare AI and XAI, making it necessary and timely to conduct an up-to-date review that captures these developments and provides actionable insights.

Therefore, this study aims to provide a comprehensive review of XAI in healthcare. Specifically, the study examines the foundational concepts of XAI, explores its diverse applications in the healthcare sector, identifies the key challenges it faces, and highlights possible solutions for its effective integration into medical practice. This study is significant given the growing importance of transparency and interpretability in the ethical deployment of AI technologies in healthcare settings.

The remainder of this paper is structured as follows: Section 2 reviews related works. Section 3 discusses the fundamental concepts of XAI, Section 4 outlines various XAI methods and techniques. Section 5 explores the applications of XAI in healthcare, Section 6 examines responsible AI in healthcare, and Section 7 identifies the challenges and opportunities in implementing XAI within clinical settings. Section 8 provides a comprehensive discussion on future research directions, identifying areas of potential growth and innovation. Section 9 concludes the paper with a summary of key findings.

2. Related Works

The application of AI in healthcare has been extensively explored in recent years, with different studies focusing on various aspects of its implementation and impact. For instance, Topol [12] provides an in-depth analysis of AI's capabilities in clinical settings, including improvements in diagnostic processes and patient care. Similarly, Malik et al. [13] presented an overview of AI applications in medicine, detailing various AI models and their effectiveness in predicting and managing diseases. These studies showed the significant advancements AI brings to healthcare.

Meanwhile, XAI is a crucial subfield of AI which addresses the need for transparency and interpretability in AI models. Arrieta et al. [20] provided a comprehensive review of XAI techniques, categorizing them into model-specific and model-agnostic approaches and highlighting the importance of XAI in enabling trust and accountability in AI models. Similarly, Samek et al. [21] focused on the theoretical foundations of XAI, presenting various methods to make AI models interpretable and demonstrating the need for these methods to gain acceptance in high-stakes domains such as healthcare.

Adadi and Berrada [22] examined the potential of XAI in various sectors, including healthcare, identifying the key challenges in making AI models interpretable and proposing potential solutions to address these challenges. The study provides valuable insights into the importance of XAI. Furthermore, a critical aspect of AI that has gained attention recently is the concept of responsible AI, which encompasses ethical considerations such as fairness, accountability, and transparency. Dignum [23] outlines the principles of responsible AI, stressing the importance of integrating ethical considerations into AI development. Similarly, Barocas et al. [24] discussed fairness in AI, proposing methods to detect and mitigate biases in ML models.

Furthermore, Holzinger et al. [25] discuss the need for human-in-the-loop (HITL) approaches in AI, particularly in healthcare, arguing that for AI systems to be trusted and widely adopted, they must not only be accurate but also provide explanations that are understandable to users. Their work highlights the theoretical foundations of XAI and the practical need for HITL approaches to ensure that AI systems can be effectively integrated into clinical practice.

Meanwhile, most existing reviews and surveys on AI and XAI provide a broad overview of the field, focusing on specific applications or theoretical developments. For example, Jiang et al. [11] reviewed the application of AI in healthcare broadly, without discussing the specific challenges of explainability. Similarly, Miotto et al. [26] focused on deep learning in healthcare, offering insights into its potential but not addressing the interpretability challenges associated with these models.

 doi:10.20944/preprints202408.1702.v1

Therefore, this review aims to provide a detailed and comprehensive analysis of XAI in healthcare, covering key areas such as foundational concepts, diverse applications, challenges, and opportunities for future research. By focusing on these aspects, this review aims to bridge the gap in the literature and offer actionable insights for researchers, practitioners, and policymakers in the field of healthcare AI.

3. Overview of Explainable AI

XAI is a domain within AI focused on creating models whose decisions can be understood and interpreted by humans. The primary goal of XAI is to make the internal mechanics of AI systems transparent and their outputs explainable [27,28]. This is essential in fields like healthcare, where understanding the reasoning behind AI decisions can significantly impact patient outcomes and trust in the technology. The main goals of XAI include:

3.1. Transparency

Transparency involves making the AI decision-making process clear and understandable [29, 30]. This means that the inner workings of an AI model, such as the data it uses, the features it considers, and the logic it follows to reach a decision, should be visible to and interpretable by humans. Transparency is crucial for identifying potential biases, ensuring fairness, and gaining trust. Mathematically, transparency can be represented by ensuring the model f is such that for any input x, the decision process $f(x)$ can be decomposed into understandable components:

$$f(x) = \sum_{i=1}^{n} w_i \cdot x_i + b, \tag{1}$$

where w_i are the weights and x_i are the input features, providing a linear combination that is easily interpretable [31].

3.2. Interpretability

Interpretability refers to the extent to which a human can understand the cause of a decision made by an AI model. This involves providing explanations that are comprehensible to humans, often through simplified models or visualizations that capture the behavior of the complex model [32]. Interpretability can be achieved through various methods, such as surrogate models, partial dependence plots, and feature importance scores.

3.3. Trustworthiness

Trustworthiness involves building user trust through understandable and verifiable AI decisions [20,32]. Trust is essential for the widespread adoption of AI technologies, particularly in critical fields like healthcare [33]. Trustworthiness can be achieved by ensuring that AI systems are transparent, interpretable, and robust. This includes providing clear documentation of the model's decision-making process, using robust validation techniques to ensure the model's reliability, and continuously monitoring the model's performance to detect and address any issues promptly.

For instance, incorporating human-in-the-loop approaches, where human experts interact with AI systems to validate and refine their outputs, can enhance trustworthiness. Involving clinicians in the decision-making process ensures AI systems can benefit from expert knowledge and feedback, leading to more accurate and reliable outcomes [34]. Additionally, frameworks that track and log AI decisions, providing an audit trail, can help users review and understand the rationale behind AI-driven decisions, further enhancing trust.

 doi:10.20944/preprints202408.1702.v1

3.4. Accountability

Accountability involves enabling users to hold AI systems accountable for their decisions. This means that AI systems should provide enough information to allow users to understand, challenge, and, if necessary, rectify the decisions made by the AI [35]. Accountability is essential for ensuring ethical AI deployment. This can be supported by frameworks that track and log AI decisions, providing an audit trail that users can review. For example, in regression models, accountability can be enhanced by providing confidence intervals for predictions:

$$\hat{y} = \beta_0 + \beta_1 x + \epsilon, \tag{2}$$

where $\hat{y}$ is the predicted value, β_0 and β_1 are coefficients, and ϵ represents the error term. The confidence interval gives users an idea of the uncertainty in the prediction, helping them to hold the model accountable for its predictions [36]. Furthermore, the key terms and concepts associated with explainable AI are tabulated in Table 1.

 doi:10.20944/preprints202408.1702.v1

Table 1. Key Terms and Concepts in Explainable AI.

Term	Description
Interpretable ML	An interpretable model is one where a user can see and understand how inputs are mathematically mapped to outputs.
Black-box problem	The challenge in AI where the internal workings of an AI model are not visible or understandable to the user, often leading to a lack of trust and transparency.
XAI	A set of processes and methods that allow human users to comprehend and trust the results and outputs created by ML algorithms [6,28].
Responsible AI	AI that takes into account societal values, moral, and ethical considerations, focusing on accountability, responsibility, and transparency [37].
Fairness in AI	Ensuring that AI systems make decisions impartially, without bias towards any group.
Accountability in AI	The obligation of AI systems to provide explanations for their decisions, enabling users to understand, challenge, and rectify AI-driven outcomes [38].
Transparency in AI	Making the decision-making processes of AI systems visible and understandable to users, ensuring clarity in how AI systems operate and make decisions [33].
Trustworthy AI	AI systems that are reliable, robust, and have a high degree of integrity, gaining user trust through transparency, fairness, and accountability.
Causability	The ability to provide causal explanations for AI decisions, moving beyond mere correlations to understand the underlying causes of outcomes [39].
Human-in-the-loop	A model in AI where human judgment is integrated into the AI system's decision-making process to enhance accuracy, fairness, and accountability [40,41].
Cognitive Bias in AI	The phenomenon where AI systems may inadvertently learn and perpetuate human biases present in the training data, leading to biased outcomes [42].
Ethical AI	The practice of designing and deploying AI systems in ways that are aligned with ethical principles, such as fairness, accountability, and transparency. [43,44]
Data Privacy	The protection of personal data used in AI systems, ensuring that sensitive information is handled securely and ethically [45,46].

4. XAI Techniques and Methods

Several techniques and methods have been developed to achieve the goals of XAI. These methods can be broadly classified into two categories: model-specific and model-agnostic techniques.

 doi:10.20944/preprints202408.1702.v1

4.1. Model-Specific Techniques

Model-specific techniques are tailored to particular types of models. These model-specific techniques enhance the interpretability of AI models by providing clear, understandable structures and visualizations that help users comprehend how decisions are made. For instance:

4.1.1. Decision Trees and Rule-Based Systems

Decision trees and rule-based systems are inherently interpretable because they follow a clear structure of decisions and rules that can be easily understood [47,48]. Each decision in a decision tree represents a choice based on a specific feature, making it straightforward to trace the path from the root to a leaf node (final decision) [49]. For example, a decision tree model f can be represented as a set of nested if-then rules:

$$f(x) = \sum_{i=1}^{n} \text{if } (x_i < \theta_i) \text{ then } a_i \text{ else } b_i, \tag{3}$$

where x_i are the input features, θ_i are the threshold values, and a_i and b_i are the decisions or outputs at each node. This structure allows users to understand how the model arrives at a specific decision by following the path dictated by the feature values [49]. The structure of a typical decision tree is shown in Figure 1.

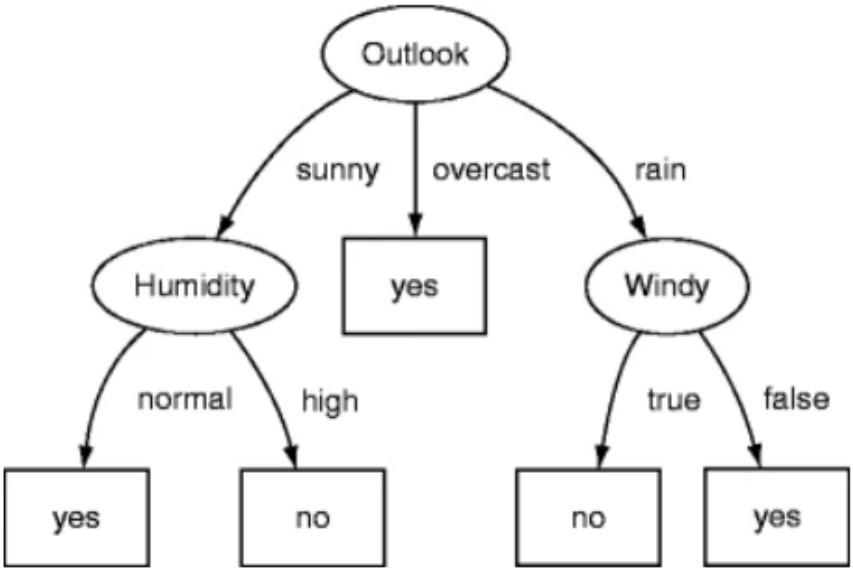

Figure 1. Example of a Decision Tree [50].

4.1.2. Attention Mechanisms in Neural Networks

Attention mechanisms in neural networks provide insights into which parts of the input data the model is focusing on when making a decision, thus offering some level of interpretability [51,52]. Attention mechanisms assign different weights to different parts of the input, highlighting their relative importance in the final decision [53]. The attention mechanism can be represented mathematically as:

$$\alpha_i = \frac{\exp(e_i)}{\sum_{j=1}^{n} \exp(e_j)}, \tag{4}$$

where α_i is the attention weight for the i-th input, and e_i is the alignment score between the input x_i and the model's internal state. The final output of the attention mechanism is a weighted sum of the

 doi:10.20944/preprints202408.1702.v1

input features. This weighted sum c allows users to visualize and interpret which input features are most influential in the model's decision-making process [54]. It is represented mathematically as:

$$c = \sum_{i=1}^{n} \alpha_i x_i. \tag{5}$$

4.1.3. Convolutional Neural Networks

In Convolutional Neural Networks (CNNs), interpretability can be enhanced through techniques like Gradient-weighted Class Activation Mapping (Grad-CAM). Grad-CAM provides visual explanations for decisions made by CNNs by highlighting the regions of an input image that are most relevant to the prediction [55]. The Grad-CAM heatmap $L^c_{\text{Grad-CAM}}$ for a class c is calculated as:

$$L^c_{\text{Grad-CAM}} = \text{ReLU}\left(\sum_k \alpha_k^c A^k\right), \tag{6}$$

where α_k^c is the importance weight for the k-th feature map A^k, and ReLU is the rectified linear unit activation function. This heatmap overlays the original image, showing which regions contributed most to the model's decision [55].

4.1.4. Bayesian Networks

Bayesian networks are probabilistic graphical models that represent a set of variables and their conditional dependencies via a directed acyclic graph (DAG) [56]. They provide interpretability by visualizing the probabilistic relationships between variables. The joint probability distribution P over a set of variables X can be decomposed as:

$$P(X) = \prod_{i=1}^{n} P(X_i|\text{Parents}(X_i)), \tag{7}$$

where $\text{Parents}(X_i)$ denotes the set of parent nodes for X_i in the DAG. This decomposition allows users to understand how each variable influences others and contributes to the overall model's predictions [57].

4.2. Model-Agnostic Techniques

Model-agnostic techniques can be applied to any AI model, irrespective of its underlying architecture. These techniques provide flexibility and can be used to interpret complex models without requiring changes to the model itself. Prominent model-agnostic techniques include:

4.2.1. SHapley Additive exPlanations

SHapley Additive exPlanations (SHAP) values are a method based on cooperative game theory that provides a unified measure of feature importance [58]. This method assigns an importance value to each feature by computing the Shapley value, which represents the average contribution of a feature across all possible combinations of features. The Shapley value ϕ_i for a feature i is calculated as:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} \left[f(S \cup \{i\}) - f(S)\right], \tag{8}$$

where S is any subset of features not containing i, and $f(S)$ is the prediction from the model considering only the features in S [59]. This equation ensures that the sum of the Shapley values equals the difference between the actual prediction and the average prediction, thus providing a fair distribution of feature importance. SHAP values have several desirable properties that make them a powerful tool

 doi:10.20944/preprints202408.1702.v1

for interpreting ML models. Firstly, they provide consistency, meaning that if a model's prediction for a particular instance increases due to a change in a feature value, the Shapley value for that feature will also increase. Secondly, SHAP values offer both global and local interpretability. Global interpretability refers to understanding the overall importance of each feature across the entire dataset, while local interpretability focuses on understanding how features influence individual predictions [60]. For example, in a healthcare setting, SHAP values can be used to interpret a model predicting the risk of heart disease. By examining the Shapley values, clinicians can identify which features (e.g., age, cholesterol levels, blood pressure) are most influential in the model's predictions for individual patients and across the patient population.

Meanwhile, SHAP values can be visualized using various plots to aid in interpretability. The most common visualizations include summary plots, dependence plots, and force plots. Summary plots provide a high-level overview of feature importance across the dataset, highlighting the distribution of Shapley values for each feature. Dependence plots show the relationship between a feature's value and its Shapley value, indicating how changes in the feature value impact the model's prediction. Force plots offer a detailed view of individual predictions, illustrating how each feature contributes to the final prediction [61]. An example architecture for integrating SHAP with an ML model is shown in Figure 2.

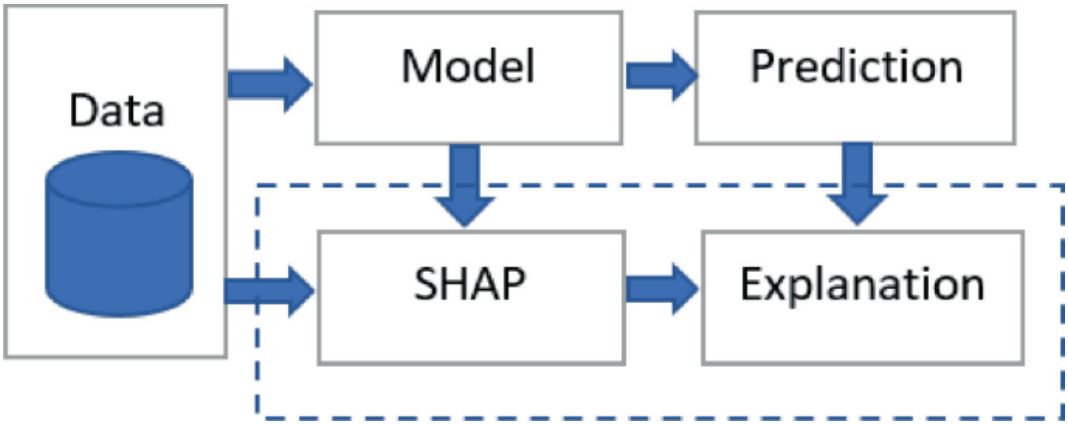

Figure 2. SHAP-ML Model Architecture [**?**]

4.2.2. Local Interpretable Model-agnostic Explanations

Local Interpretable Model-agnostic Explanations (LIME) is a technique designed to explain individual predictions of any black-box model by approximating it locally with an interpretable model [62]. LIME operates by perturbing the data around the instance of interest, generating a dataset of perturbed samples, and then training an interpretable model (often a linear regression or decision tree) on this perturbed dataset. The weights in this interpretable model are used to explain the prediction of the original model. This approach allows for an understanding of the complex model's behavior in the vicinity of the specific instance being explained. Given a black-box model f and an instance x, LIME constructs a new dataset Z consisting of perturbed samples of x and their corresponding predictions from f. A weighted linear model g is then trained on Z, where the weights are based on the proximity of the perturbed samples to x. The explanation is provided by the coefficients of the linear model g:

$$g(z) = \arg\min_{g \in G} \sum_{z_i \in Z} \pi_x(z_i)(f(z_i) - g(z_i))^2 + \Omega(g), \tag{9}$$

where $\pi_x(z_i)$ is a proximity measure between x and z_i, and $\Omega(g)$ is a complexity measure for g [62,63]. Algorithm 1 summarizes the LIME process:

 doi:10.20944/preprints202408.1702.v1

Algorithm 1 LIME Process

Require: Black-box model f, instance x, number of perturbations N
1: Generate a new dataset Z by perturbing x N times
2: **for** each perturbed instance $z_i \in Z$ **do**
3: Obtain prediction $f(z_i)$ from the black-box model
4: Compute the proximity measure $\pi_x(z_i)$ between x and z_i
5: **end for**
6: Train a weighted linear model g on Z, using $\pi_x(z_i)$ as weights
7: Use the coefficients of g to explain the prediction for x
8: **return** Explanation of x based on g

By focusing on the local behavior of the model around a specific instance, LIME provides an understandable approximation that can highlight which features are driving a particular prediction. LIME's utility in healthcare is vast. For instance, in the context of predicting patient outcomes, LIME can help clinicians understand which features (e.g., patient age, lab results, medical history) are influencing the model's prediction for a specific patient. This local explanation is crucial in making the model's decision-making process transparent and comprehensible to healthcare professionals who may not have a deep understanding of ML models.

4.2.3. Partial Dependence Plots

Partial Dependence Plots (PDPs) show the relationship between a subset of features and the predicted outcome of a machine learning model [64,65]. The partial dependence function for a feature x_j is defined as:

$$\hat{f}_{x_j}(x_j) = \frac{1}{n} \sum_{i=1}^{n} \hat{f}(x_j, x_{iC}), \tag{10}$$

where x_{iC} represents all features except x_j, and $\hat{f}$ is the prediction function. PDPs provide insight into the effect of a feature on the prediction while averaging out the effects of other features. This method helps to visualize the marginal effect of a feature on the predicted outcome, assuming that the effect of other features remains constant. For example, in a healthcare setting, PDPs can be used to understand how a single biomarker influences the risk prediction of a disease, independent of other biomarkers. This visualization aids clinicians in interpreting the importance and influence of specific features, thereby enhancing the transparency and trust in the model's predictions [66].

4.2.4. Individual Conditional Expectation Plots

Individual Conditional Expectation (ICE) plots are similar to PDPs but show the dependency of the prediction on a feature for each instance separately rather than averaging [67]. For an instance i, the ICE curve for a feature x_j is given by:

$$\hat{f}_{x_j}^{(i)}(x_j) = \hat{f}(x_j, x_{iC}). \tag{11}$$

ICE plots provide a more granular view of feature effects, revealing heterogeneity in the model's behavior across different instances [68]. This granular view is particularly useful for detecting interactions and non-linear relationships between features and the outcome. In the context of healthcare, ICE plots can show how different patients respond to varying levels of a particular treatment, thereby highlighting the variability in treatment effectiveness across the patient population.

5. Applications of Explainable AI in Healthcare

XAI significantly enhances various aspects of healthcare by improving diagnostic accuracy, treatment personalization, clinical decision support, medical imaging, and remote diagnostics. By making AI models transparent and interpretable, XAI builds trust and reliability in AI-driven healthcare systems.

 doi:10.20944/preprints202408.1702.v1

5.1. Diagnostic Tools and Clinical Decision Support Systems

Integrating XAI into diagnostic tools and Clinical Decision Support Systems (CDSS) is crucial for enhancing the interpretability and trustworthiness of AI models. In oncology, cardiovascular diagnostics, and neurological disorders, XAI techniques like SHAP values, LIME, and attention mechanisms help explain AI model predictions, thus improving the accuracy and transparency of diagnostics [69–71]. Recent studies, such as that by Lundberg et al. [72], have shown how SHAP values can be used to interpret predictions of ML models for diagnosing pneumonia from chest X-rays. Similarly, studies have demonstrated the effectiveness of XAI techniques in predicting diabetic retinopathy from retinal images, using saliency maps to highlight the critical areas influencing the model's predictions [73].

XAI is also instrumental in CDSS, where it elucidates the factors influencing diagnostic and treatment recommendations. For example, in sepsis prediction for ICU patients, Li et al. [74] used LIME interpret model predictions, enhancing the transparency and trustworthiness of the CDSS. Similarly, Bedoya et al. [15] SHAP values have been employed in CDSS for predicting hospital readmissions, helping healthcare providers tailor care plans to individual patients. Another study by Tonekaboni et al. [75] explored the use of attention mechanisms within a CDSS for diagnosing acute kidney injury (AKI), highlighting the most relevant data points and thereby improving diagnostic accuracy.

Additionally, the integration of XAI in models used for detecting arrhythmias from ECG data has shown significant promise. A study by Bento et al. [76] demonstrated how XAI techniques could enhance the transparency of deep learning models, providing cardiologists with clear visual and statistical explanations of abnormal heart rhythms, thus supporting the integration of these advanced diagnostic tools into clinical practice.

5.2. Personalized Medicine

XAI plays a crucial role in personalized medicine by making the outputs of ML models more interpretable. These models analyze genetic information, medical history, and lifestyle factors to recommend individualized treatment plans. XAI techniques like SHAP values and LIME provide clear explanations for these recommendations, ensuring that treatments are tailored to each patient's needs and improving both outcomes and adherence. For instance, in the management of type 2 diabetes, SHAP values have been used to interpret ML models that predict the effectiveness of different medications, allowing clinicians to understand how factors such as age, weight, and blood sugar levels influence treatment recommendations. Another significant application is in oncology, where XAI has been employed to tailor chemotherapy treatments. Techniques like LIME have been used to explain the genetic markers and clinical features influencing treatment plans, improving the interpretability and trustworthiness of AI-driven recommendations [77].

Additionally, XAI has shown potential in managing cardiovascular diseases. A study by Ahmad et al. [78] demonstrated the use of XAI to interpret ML models that predict the risk of adverse cardiac events, such as heart attacks. By providing clear explanations for these risk predictions, XAI helps cardiologists develop personalized prevention strategies, thereby enhancing patient outcomes. Furthermore, XAI has been applied in the treatment of rare genetic disorders. A study by Ravindran et al. [79] used XAI to interpret deep learning models predicting the efficacy of gene therapies, enabling geneticists to understand and validate the AI's recommendations.

5.3. Medical Imaging

In medical imaging, XAI enhances the interpretability of ML models used for image analysis, crucial for detecting and diagnosing conditions from X-rays, MRIs, and CT scans. Techniques like attention maps and Grad-CAM highlight the image regions most influential in the model's decision, providing radiologists with visual explanations that increase confidence in AI findings [80].

 doi:10.20944/preprints202408.1702.v1

Recent studies have highlighted the effectiveness of XAI in various medical imaging applications. For instance, McKinney et al. [81] applied XAI techniques to deep learning models used for breast cancer screening from mammograms. The use of attention maps allowed radiologists to see which parts of the mammogram the model focused on, improving diagnostic accuracy and reducing false positives and negatives. In another study, DeGrave et al. [82] investigated the use of XAI in the diagnosis of COVID-19 from chest X-rays, employing saliency maps to visualize the lung regions crucial for detecting the infection.

Furthermore, a study by Baumgartner et al. [83] explored the application of Grad-CAM in MRI-based brain tumor classification, demonstrating how specific brain regions contributing to the AI's predictions could be visualized, thereby enhancing diagnostic confidence and potentially leading to more accurate treatment plans. In ophthalmology, XAI has been used to interpret ML models for diagnosing diabetic retinopathy from retinal images, with studies like Gargeya and Leng [84] showing how heatmaps can help ophthalmologists understand and trust AI-driven diagnoses.

5.4. Remote Diagnostics and Telemedicine

XAI also improves the accessibility and accuracy of remote diagnostics and telemedicine, particularly in underserved areas. By ensuring that AI-driven diagnostic suggestions are accompanied by clear explanations, XAI helps remote healthcare providers understand and trust AI recommendations, enhancing care quality where access to specialists is limited [85]. For example, in dermatology, Hauser et al. [86] XAI has been used to interpret ML models diagnosing skin conditions via mobile devices, with heatmaps highlighting key regions of interest. This approach supports healthcare workers in remote locations, ensuring accurate and timely diagnoses [87]. Similarly, XAI enhances the trustworthiness of telemedicine platforms for diagnosing eye diseases from retinal images, as demonstrated by Arcadu et al. [88], improving the quality of care in underserved regions.

 doi:10.20944/preprints202408.1702.v1

Table 2. Summary of XAI Applications in Healthcare.

Application	Specific Use Case	Description	References
Diagnostic Tools and CDSS	Oncology	Identifying cancerous lesions using medical imaging data	[69]
	Cardiovascular Diseases	Detecting patterns in ECGs indicative of heart diseases	[70]
	Pneumonia Detection	Using SHAP values for pneumonia diagnosis from chest X-rays	[72]
	Diabetic Retinopathy	Interpreting retinal images to predict diabetic retinopathy	[73]
	Neurological Disorders	Diagnosing Alzheimer's using attention mechanisms on MRI scans	[71]
	Arrhythmia Detection	Enhancing transparency in deep learning models for arrhythmia diagnosis	[89]
Personalized Medicine	Type 2 Diabetes	Personalized treatment recommendations based on electronic health records	[78]
	Oncology	Tailoring chemotherapy treatments using genetic profiles and clinical data	[77]
	Cardiovascular Diseases	Predicting risk of cardiac events and tailoring prevention strategies	[79]
	Gene Therapies	Recommending gene editing techniques based on genomic data	[79]
Medical Imaging	Breast Cancer	Applying attention maps in deep learning models for mammogram analysis	[81]
	COVID-19 Detection	Using saliency maps to diagnose COVID-19 from chest X-rays	[82]
	Brain Tumors	Visualizing brain regions in MRI scans with Grad-CAM for tumor classification	[83]
	Diabetic Retinopathy	Highlighting retinal areas in AI predictions for diabetic retinopathy diagnosis	[84]
Remote Diagnostics and Telemedicine	Respiratory Diseases	Integrating XAI in telemedicine platforms for respiratory disease diagnosis	[26]
	Dermatology	Employing XAI for diagnosing skin conditions via mobile devices	[86]
	Ophthalmology	Interpreting AI models in telemedicine for eye disease diagnosis from retinal images	[88]

6. Responsible AI in Healthcare

The deployment of AI in healthcare necessitates a focus on responsible AI, which encompasses ethical considerations, accountability, transparency, and fairness. Responsible AI aims to ensure that AI systems are developed and used in ways that respect human rights, promote fairness, and enhance societal well-being. There are different aspects of responsible AI:

6.1. Ethical Considerations

Ethical considerations are paramount in the development and deployment of AI systems in healthcare. AI systems must be designed to uphold patient privacy and data security, ensuring that sensitive health information is protected throughout the data processing lifecycle [37]. This involves

 doi:10.20944/preprints202408.1702.v1

implementing robust data encryption methods, anonymizing patient data, and ensuring compliance with regulations such as the Health Insurance Portability and Accountability Act (HIPAA) and the General Data Protection Regulation (GDPR) [90]. Additionally, ethical AI practices must include obtaining informed consent from patients regarding the use of their data and providing information on how their data will be used and protected.

6.2. Accountability and Transparency

Accountability in AI involves the ability to explain and justify AI-driven decisions, enabling users to understand and challenge these decisions if necessary. In healthcare, accountability is critical as AI systems are often involved in high-stakes decisions that can significantly impact patient outcomes. To achieve accountability, AI models must be designed to provide clear, understandable explanations for their predictions and decisions. This requires incorporating XAI techniques that make the decision-making process transparent [38]. Transparency in AI systems involves making the decision-making processes visible and understandable to users [91]. This can be achieved through various XAI techniques such as LIME and SHAP, which shows how AI models arrive at their conclusions. Ensuring transparency and accountability in AI systems builds trust among healthcare professionals and patients. When users understand how AI models make decisions, they are more likely to trust and adopt these technologies in clinical practice. This trust is essential for the successful integration of AI into healthcare workflows, ultimately leading to better patient outcomes and more effective use of AI technologies.

6.3. Fairness and Bias Mitigation

Fairness in AI refers to the impartiality and equity in AI decision-making processes [92]. AI systems must be rigorously tested for biases that could lead to discriminatory outcomes. Bias can occur at various stages of the AI lifecycle, including data collection, model training, and deployment [93]. Techniques such as re-sampling the training data, adjusting model parameters, and incorporating fairness constraints during model training are essential to mitigate these biases [42]. Furthermore, to ensure fairness, it is important to use diverse and representative datasets that reflect the population the AI system will serve. This prevents the model from learning and perpetuating existing biases. Additionally, bias detection tools can be employed to identify and address any unfair patterns in the AI system's decisions. Also, continuous monitoring and auditing of AI systems are crucial for maintaining fairness over time [94]. Fair AI systems contribute to building trust among users and ensuring the widespread adoption of AI technologies in healthcare. When AI systems are fair, they are more likely to be accepted and trusted by healthcare professionals and patients. This trust is vital for the effective deployment of AI technologies, which can lead to more equitable healthcare outcomes and improved patient care.

6.4. Human-in-the-loop Approaches

Human-in-the-loop (HITL) approaches integrate human judgment into AI decision-making processes, enhancing the accuracy, fairness, and accountability of AI systems [40]. HITL models ensure that human oversight is maintained, particularly in critical healthcare decisions, thereby increasing the trustworthiness of AI systems. Incorporating HITL approaches involves designing AI systems that allow for human intervention and feedback. This can include interactive interfaces where clinicians can review and modify AI-generated recommendations before final decisions are made. Such systems leverage the strengths of both AI and human expertise, leading to more accurate and reliable outcomes [95]. Furthermore, HITL approaches help to address the limitations of AI models by allowing humans to correct errors and biases that the AI might not detect. This collaborative approach ensures that AI systems are continually improved and aligned with clinical standards and ethical guidelines. By integrating human judgment with AI capabilities, HITL approaches enhance the overall effectiveness and trustworthiness of AI in healthcare.

 doi:10.20944/preprints202408.1702.v1

7. Challenges and Opportunities in Implementing XAI in Healthcare

Despite the significant advancements in XAI, several challenges persist in its implementation within healthcare settings, such as:

7.1. Integration into Clinical Workflows

Integrating XAI methods seamlessly into existing clinical workflows is a major challenge. AI models must provide explanations that are both accurate and understandable to healthcare professionals who may not have technical backgrounds. Developing user-friendly interfaces that clearly present AI-generated explanations is crucial, as interactive visualization tools can assist clinicians in interpreting complex AI outputs, thereby enhancing their decision-making processes. Ensuring that XAI systems are interoperable with various healthcare IT systems, such as electronic health records (EHRs), is also vital. Standardized data formats and protocols can facilitate this interoperability, allowing different systems to communicate and share information seamlessly [96]. Furthermore, training and educating healthcare professionals is essential, equipping them with the necessary skills to understand and interpret AI-generated explanations and enabling a culture of trust and collaboration [12].

7.2. Regulatory Compliance

Regulatory compliance remains a significant hurdle in the implementation of XAI in healthcare. AI systems must adhere to strict regulations such as the Health Insurance Portability and Accountability Act in the United States and GDPR in Europe. These regulations emphasize the need for transparency and accountability in AI-driven decision-making processes. Therefore, XAI methods must be designed to meet these regulatory standards, providing clear documentation and audit trails for their decisions [90]. Compliance involves robust data encryption, anonymization of patient data, and ensuring transparency and auditability in data processing activities. Adhering to these regulations not only builds ethical and trustworthy systems but also gains the trust of healthcare professionals and patients. AI systems that comply with legal and ethical standards are more likely to be accepted and integrated into clinical practice, ensuring responsible AI use in healthcare settings [97].

7.3. Bias and Fairness

Biases can be introduced during various stages of the AI lifecycle, such as data collection and model training. For instance, an AI system trained on a dataset lacking diversity may perform poorly on underrepresented populations, worsening healthcare disparities. XAI techniques, such as feature importance and counterfactual explanations, can highlight biased patterns and facilitate the development of more equitable models [98]. However, identifying and quantifying bias in complex models can be difficult, and existing XAI methods may not always provide sufficient granularity to detect subtle biases. Additionally, mitigating bias without compromising the model's performance poses a significant challenge. Ensuring robust and generalizable bias detection and mitigation techniques across different healthcare settings is critical [99,100].

7.4. Interpretability vs. Accuracy

Balancing interpretability and accuracy is challenging in XAI. Highly interpretable models, such as linear regression and decision trees, often have lower accuracy compared to complex models like deep neural networks [101]. This trade-off can limit the effectiveness of XAI in clinical settings where both accuracy and interpretability are crucial. The challenge is to find an optimal balance where the model remains sufficiently interpretable without significantly compromising its accuracy. Complex models, often referred to as "black boxes" due to their intricate internal structures, pose difficulties in interpretation. Simplifying these models can lead to a loss of critical predictive power, thus affecting clinical outcomes [32].

 doi:10.20944/preprints202408.1702.v1

Opportunities to address this challenge include developing hybrid models that combine interpretable components with complex models. Techniques such as surrogate models, where an interpretable model approximates the behavior of a complex model, can provide explanations while maintaining accuracy. Advanced XAI methods like SHAP and LIME offer detailed insights into model predictions without significantly reducing accuracy. Ongoing research into more transparent architectures for complex models, such as inherently interpretable neural networks, holds promise for balancing interpretability and accuracy [33,102].

7.5. Long-term Impact on Patient Outcomes

The long-term impact of XAI on patient outcomes and the overall healthcare system requires further exploration. Longitudinal studies are needed to assess how the integration of XAI affects clinical practices over time, including its influence on patient trust, treatment adherence, and health outcomes. Understanding these impact could provide valuable insights into the effectiveness of XAI and guides future improvements in AI-driven healthcare solutions.

8. Discussion and Future Research Directions

This study has identified that one of the main benefits of XAI in healthcare is enhancing the transparency and trust in AI-driven decisions. Clinicians require clear and understandable explanations for AI predictions to make informed decisions, especially in critical scenarios such as diagnosing diseases or developing treatment plans. Techniques like SHAP and LIME have proven effective in providing these explanations, bridging the gap between complex AI models and clinical applicability [103]. Despite these advancements, there is an ongoing need for more user-friendly interfaces that present AI explanations in an easily understandable format for healthcare professionals.

Another major challenge is balancing interpretability and accuracy. Simpler, interpretable models may lack the predictive power of more complex algorithms, limiting their utility in clinical settings. Research into hybrid models that combine interpretable and high-accuracy components offers a promising solution. Achieving an optimal balance between interpretability and accuracy remains a significant challenge that needs further exploration [32,102].

Ensuring regulatory compliance is another critical aspect. AI systems in healthcare must adhere to strict regulations such as HIPAA and GDPR to ensure data privacy and security. XAI methods must be designed to meet these regulatory standards, including robust data encryption and anonymization techniques. The study emphasizes the importance of maintaining comprehensive audit trails for AI decisions to ensure accountability and transparency, which are essential for gaining trust from healthcare providers and patients alike [90,104].

Ethical considerations are paramount in the deployment of XAI in healthcare. The potential for biases in AI models to lead to discriminatory outcomes and even worsen existing healthcare disparities is a significant concern. XAI can help identify and mitigate these biases, but ensuring that these methods are effective across diverse populations is an ongoing challenge. Developing comprehensive frameworks for bias detection and mitigation, involving collaboration between AI researchers, healthcare professionals, and ethicists, is crucial for promoting equity in healthcare delivery [42,99].

Therefore, future research directions in XAI should focus on the following:

- **Developing inherently interpretable models**: Future research should prioritize the creation of models that are transparent by design, reducing the reliance on post-hoc explanation techniques. Inherently interpretable models can provide direct insights into their decision-making processes, making them more trustworthy and easier to integrate into clinical workflows.
- **Integrating causal inference techniques**: Combining causal inference with XAI can provide deeper insights into how different variables influence outcomes, which is particularly valuable in clinical settings. This integration can help in understanding the causal relationships within the data, leading to more robust and reliable AI models.

 doi:10.20944/preprints202408.1702.v1

- **Advancing visualization tools**: Improved visual representations of model explanations, such as interactive dashboards and 3D visualizations, can enhance the usability of XAI tools for healthcare professionals. These advanced visualization techniques can help clinicians better understand and trust AI-driven decisions.
- **Enhancing model robustness and generalization ability**: Research should focus on developing XAI methods that are robust and generalizable across different healthcare settings and patient populations. This includes ensuring that XAI techniques can handle diverse and heterogeneous data sources, which are common in healthcare.
- **Exploring the socio-economic impact of XAI**: Research should also focus on the socio-economic impact of XAI in healthcare, including its potential to reduce healthcare disparities and improve access to quality care. Understanding these broader impacts can help in designing XAI systems that are technically sound and also socially beneficial.

By addressing these research directions, the integration of XAI in healthcare can be significantly improved, leading to more transparent, trustworthy, and effective AI-driven solutions that benefit both clinicians and patients.

9. Conclusions

This paper has provided a comprehensive review of XAI in healthcare, highlighting its potential to improve clinical decision-making, patient outcomes, and regulatory compliance. Techniques such as SHAP and LIME have proven effective in making complex AI models more interpretable and accessible to healthcare professionals, and these techniques were examined in detail. Additionally, the study explored challenges in healthcare AI and XAI, including the trade-off between interpretability and accuracy, the integration of XAI into clinical workflows, and the need for robust regulatory compliance. Challenges, opportunities, and future directions were also analyzed, contributing significantly to the literature. The study emphasized the need to focus on developing inherently interpretable models, integrating causal inference techniques, advancing visualization tools, and ensuring the ethical implications of XAI are addressed, which can lead to more transparent, trustworthy, and effective AI-driven healthcare solutions.

References

1. Lee, C.H.; Wang, C.; Fan, X.; Li, F.; Chen, C.H. Artificial intelligence-enabled digital transformation in elderly healthcare field: scoping review. *Advanced Engineering Informatics* **2023**, *55*, 101874.
2. Obaido, G.; Mienye, I.D.; Egbelowo, O.F.; Emmanuel, I.D.; Ogunleye, A.; Ogbuokiri, B.; Mienye, P.; Aruleba, K. Supervised machine learning in drug discovery and development: Algorithms, applications, challenges, and prospects. *Machine Learning with Applications* **2024**, *17*, 100576.
3. Agarwal, P.; Swami, S.; Malhotra, S.K. Artificial intelligence adoption in the post COVID-19 new-normal and role of smart technologies in transforming business: a review. *Journal of Science and Technology Policy Management* **2024**, *15*, 506–529.
4. Uysal, İ.; Kose, U. Explainability and the Role of Digital Twins in Personalized Medicine and Healthcare Optimization. In *Explainable Artificial Intelligence (XAI) in Healthcare*; CRC Press, 2024; pp. 141–156.
5. Lesley, U.; Kuratomi Hernández, A. Improving XAI Explanations for Clinical Decision-Making–Physicians' Perspective on Local Explanations in Healthcare. International Conference on Artificial Intelligence in Medicine. Springer, 2024, pp. 296–312.
6. Gaur, L.; Gaur, D. Explainable Artificial Intelligence (XAI) on Neurogenerative Diseases. In *AI and Neuro-Degenerative Diseases: Insights and Solutions*; Springer, 2024; pp. 63–72.
7. Rahman, A.; Debnath, T.; Kundu, D.; Khan, M.S.I.; Aishi, A.A.; Sazzad, S.; Sayduzzaman, M.; Band, S.S. Machine learning and deep learning-based approach in smart healthcare: Recent advances, applications, challenges and opportunities. *AIMS Public Health* **2024**, *11*, 58.
8. Hulsen, T. Explainable artificial intelligence (XAI): concepts and challenges in healthcare. *AI* **2023**, *4*, 652–666.

 doi:10.20944/preprints202408.1702.v1

9. Alowais, S.A.; Alghamdi, S.S.; Alsuhebany, N.; Alqahtani, T.; Alshaya, A.I.; Almohareb, S.N.; Aldairem, A.; Alrashed, M.; Bin Saleh, K.; Badreldin, H.A.; others. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. *BMC medical education* **2023**, *23*, 689.
10. Borys, K.; Schmitt, Y.A.; Nauta, M.; Seifert, C.; Krämer, N.; Friedrich, C.M.; Nensa, F. Explainable AI in medical imaging: An overview for clinical practitioners–Beyond saliency-based XAI approaches. *European journal of radiology* **2023**, *162*, 110786.
11. Jiang, F.; Jiang, Y.; Zhi, H.; Dong, Y.; Li, H.; Ma, S.; Wang, Y.; Dong, Q.; Shen, H.; Wang, Y. Artificial intelligence in healthcare: past, present and future. *Stroke and vascular neurology* **2017**, *2*.
12. Topol, E. *Deep medicine: how artificial intelligence can make healthcare human again*; Hachette UK, 2019.
13. Malik, P.; Pathania, M.; Rathaur, V.K.; others. Overview of artificial intelligence in medicine. *Journal of family medicine and primary care* **2019**, *8*, 2328–2331.
14. Maleki Varnosfaderani, S.; Forouzanfar, M. The role of AI in hospitals and clinics: transforming healthcare in the 21st century. *Bioengineering* **2024**, *11*, 337.
15. Bedoya, J.C.L.; Castro, J.L.A. Explainability analysis in predictive models based on machine learning techniques on the risk of hospital readmissions. *Health and Technology* **2024**, *14*, 93–108.
16. Uddin, S.; Khan, A.; Hossain, M.E.; Moni, M.A. Comparing different supervised machine learning algorithms for disease prediction. *BMC medical informatics and decision making* **2019**, *19*, 1–16.
17. Blanco-Gonzalez, A.; Cabezon, A.; Seco-Gonzalez, A.; Conde Torres, D.; Antelo Riveiro, P.; Pineiro, A.; Garcia-Fandino, R. The Role of AI in Drug Discovery: Challenges, Opportunities, and Strategies. *Pharmaceuticals* **2023**, *16*. doi:10.3390/ph16060891.
18. Walters, W.P.; Barzilay, R. Critical assessment of AI in drug discovery. *Expert Opinion on Drug Discovery* **2021**, *16*, 937–947. doi:10.1080/17460441.2021.1915982.
19. Rogers, W.A.; Draper, H.; Carter, S.M. Evaluation of artificial intelligence clinical applications: Detailed case analyses show value of healthcare ethics approach in identifying patient care issues. *Bioethics* **2021**, *35*, 623–633. doi:10.1111/bioe.12885.
20. Arrieta, A.B.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.G.; et al.. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* **2020**, *58*, 82–115.
21. Samek, W.; Wiegand, T.; Müller, K.R. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *ITU Journal: ICT Discoveries* **2017**, *1*, 39–48.
22. Adadi, A.; Berrada, M. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE access* **2018**, *6*, 52138–52160.
23. Dignum, V. Responsible artificial intelligence: designing AI for human values. *ITU Journal: ICT Discoveries* **2017**.
24. Barocas, S.; Hardt, M.; Narayanan, A. *Fairness and machine learning: Limitations and opportunities*; MIT press, 2023.
25. Holzinger, A.; Biemann, C.; Pattichis, C.S.; Kell, D.B. What do we need to build explainable AI systems for the medical domain? *arXiv preprint arXiv:1712.09923* **2017**.
26. Miotto, R.; Wang, F.; Wang, S.; Jiang, X.; Dudley, J.T. Deep learning for healthcare: review, opportunities and challenges. *Briefings in Bioinformatics* **2018**, *19*, 1236–1246.
27. Hassija, V.; Chamola, V.; Mahapatra, A.; Singal, A.; Goel, D.; Huang, K.; Scardapane, S.; Spinelli, I.; Mahmud, M.; Hussain, A. Interpreting black-box models: a review on explainable artificial intelligence. *Cognitive Computation* **2024**, *16*, 45–74.
28. Dwivedi, R.; Dave, D.; Naik, H.; Singhal, S.; Omer, R.; Patel, P.; Qian, B.; Wen, Z.; Shah, T.; Morgan, G.; others. Explainable AI (XAI): Core ideas, techniques, and solutions. *ACM Computing Surveys* **2023**, *55*, 1–33.
29. Peters, U. Explainable AI lacks regulative reasons: why AI and human decision-making are not equally opaque. *AI and Ethics* **2023**, *3*, 963–974.
30. Albarracin, M.; Hipólito, I.; Tremblay, S.E.; Fox, J.G.; René, G.; Friston, K.; Ramstead, M.J. Designing explainable artificial intelligence with active inference: A framework for transparent introspection and decision-making. International Workshop on Active Inference. Springer, 2023, pp. 123–144.
31. Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F.; Pedreschi, D. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)* **2018**, *51*, 1–42.

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.
33. Doshi-Velez, F.; Kim, B. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* **2017**.
34. Holzinger, A. Interactive machine learning for health informatics: when do we need the human-in-the-loop? *Brain informatics* **2016**, *3*, 119–131.
35. Doshi-Velez, F.; Kortz, M.; Budish, R.; Bavitz, C.; Gershman, S.; O'Brien, D.; Scott, K.; Schieber, S.; Waldo, J.; Weinberger, D.; others. Accountability of AI under the law: The role of explanation. *arXiv preprint arXiv:1711.01134* **2017**.
36. Carvalho, D.V.; Pereira, E.M.; Cardoso, J.S. Machine learning interpretability: A survey on methods and metrics. *Electronics* **2019**, *8*, 832.
37. Okolo, C.T.; Aruleba, K.; Obaido, G. Responsible AI in Africa—Challenges and opportunities. *Responsible AI in Africa: Challenges and opportunities* **2023**, pp. 35–64.
38. Mittelstadt, B.D.; Allo, P.; Taddeo, M.; Wachter, S.; Floridi, L. The ethics of algorithms: Mapping the debate. *Big Data & Society* **2016**, *3*, 1–21.
39. Pearl, J. Theoretical impediments to machine learning with seven sparks from the causal revolution. *arXiv preprint arXiv:1801.04016* **2018**.
40. Kumar, S.; Datta, S.; Singh, V.; Datta, D.; Singh, S.K.; Sharma, R. Applications, Challenges, and Future Directions of Human-in-the-Loop Learning. *IEEE Access* **2024**.
41. Hong, S.R.; Hullman, J.; Bertini, E. Human factors in model interpretability: Industry practices, challenges, and needs. *Proceedings of the ACM on Human-Computer Interaction* **2020**, *4*, 1–26.
42. Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; Galstyan, A. A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys* **2021**, *54*, 1–35.
43. McLennan, S.; Fiske, A.; Tigard, D.; Müller, R.; Haddadin, S.; Buyx, A. Embedded ethics: a proposal for integrating ethics into the development of medical AI. *BMC Medical Ethics* **2022**, *23*. doi:10.1186/s12910-022-00746-3.
44. Adams, C.; Pente, P.; Lemermeyer, G.; Rockwell, G. Ethical principles for artificial intelligence in K-12 education. *Computers and Education: Artificial Intelligence* **2023**, *4*, 100131. doi:https://doi.org/10.1016/j.caeai.2023.100131.
45. Abouelmehdi, K.; Beni-Hessane, A.; Khaloufi, H. Big healthcare data: preserving security and privacy. *Journal of Big Data* **2018**, *5*. doi:10.1186/s40537-017-0110-7.
46. Price, II, W.N.; Cohen, I.G. Privacy in the age of medical big data. *Nature Medicine* **2019**, *25*, 37–43. doi:10.1038/s41591-018-0272-7.
47. Hong, J.S.; Lee, J.; Sim, M.K. Concise rule induction algorithm based on one-sided maximum decision tree approach. *Expert Systems with Applications* **2024**, *237*, 121365.
48. Verdasco, M.P.; García-Cuesta, E. An Interpretable Rule Creation Method for Black-Box Models based on Surrogate Trees–SRules. *arXiv preprint arXiv:2407.20070* **2024**.
49. Mienye, I.D.; Jere, N. A Survey of Decision Trees: Concepts, Algorithms, and Applications. *IEEE Access* **2024**.
50. Furnkranz, J., Decision Tree. In *Encyclopedia of Machine Learning*; Springer US: Boston, MA, 2010; chapter 15, pp. 263–267.
51. Brown, A.; Tuor, A.; Hutchinson, B.; Nichols, N. Recurrent neural network attention mechanisms for interpretable system log anomaly detection. Proceedings of the first workshop on machine learning for computing systems, 2018, pp. 1–8.
52. Dong, Y.; Su, H.; Zhu, J.; Zhang, B. Improving interpretability of deep neural networks with semantic information. Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4306–4314.
53. Chaudhari, S.; Mithal, V.; Polatkan, G.; Ramanath, R. An attentive survey of attention models. *ACM Transactions on Intelligent Systems and Technology (TIST)* **2021**, *12*, 1–32.
54. Bahdanau, D.; Cho, K.; Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* **2014**.

 doi:10.20944/preprints202408.1702.v1

55. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE international conference on computer vision, 2017, pp. 618–626.
56. Kitson, N.K.; Constantinou, A.C.; Guo, Z.; Liu, Y.; Chobtham, K. A survey of Bayesian Network structure learning. *Artificial Intelligence Review* **2023**, *56*, 8721–8814.
57. Pearl, J. *Probabilistic reasoning in intelligent systems: networks of plausible inference*; Elsevier, 2014.
58. Mienye, I.D.; Jere, N. Optimized Ensemble Learning Approach with Explainable AI for Improved Heart Disease Prediction. *Information* **2024**, *15*, 394.
59. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. *Advances in neural information processing systems* **2017**, *30*.
60. Ye, Z.; Yang, W.; Yang, Y.; Ouyang, D. Interpretable machine learning methods for in vitro pharmaceutical formulation development. *Food Frontiers* **2021**, *2*, 195–207.
61. Bifarin, O.O. Interpretable machine learning with tree-based shapley additive explanations: Application to metabolomics datasets for binary classification. *Plos one* **2023**, *18*, e0284315.
62. Zhao, X.; Huang, W.; Huang, X.; Robu, V.; Flynn, D. Baylime: Bayesian local interpretable model-agnostic explanations. Uncertainty in artificial intelligence. PMLR, 2021, pp. 887–896.
63. Ribeiro, M.T.; Singh, S.; Guestrin, C. " Why should i trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp. 1135–1144.
64. Molnar, C.; Freiesleben, T.; König, G.; Herbinger, J.; Reisinger, T.; Casalicchio, G.; Wright, M.N.; Bischl, B. Relating the partial dependence plot and permutation feature importance to the data generating process. World Conference on Explainable Artificial Intelligence. Springer, 2023, pp. 456–479.
65. Moosbauer, J.; Herbinger, J.; Casalicchio, G.; Lindauer, M.; Bischl, B. Explaining hyperparameter optimization via partial dependence plots. *Advances in Neural Information Processing Systems* **2021**, *34*, 2280–2291.
66. Peng, J.; Zou, K.; Zhou, M.; Teng, Y.; Zhu, X.; Zhang, F.; Xu, J. An explainable artificial intelligence framework for the deterioration risk prediction of hepatitis patients. *Journal of medical systems* **2021**, *45*, 61.
67. Goldstein, A.; Kapelner, A.; Bleich, J.; Pitkin, E. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *journal of Computational and Graphical Statistics* **2015**, *24*, 44–65.
68. Molnar, C.; König, G.; Bischl, B.; Casalicchio, G. Model-agnostic feature importance and effects with dependent features: a conditional subgroup approach. *Data Mining and Knowledge Discovery* **2023**, pp. 1–39.
69. Holzinger, A.; Langs, G.; Denk, H.; Zatloukal, K.; Müller, H. Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **2019**, *9*, e1312.
70. Tjoa, E.; Guan, C. A survey on explainable artificial intelligence (XAI): Toward medical XAI. *IEEE Transactions on Neural Networks and Learning Systems* **2020**, *32*, 4793–4813.
71. Thom, M.; Kreil, D.P.; Küffner, R.; Holzinger, A. Interpretable image recognition with hierarchical prototypes. *Journal of Medical Systems* **2019**, *43*, 1–13.
72. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems* **2017**, *30*, 4765–4774.
73. De Fauw, J.; Ledsam, J.R.; Romera-Paredes, B.; Nikolov, S.; Tomasev, N.; Blackwell, S.; Askham, H.; Glorot, X.; O'Donoghue, B.; Visentin, D.; others. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature Medicine* **2018**, *24*, 1342–1350.
74. Li, X.; Wu, R.; Zhao, W.; Shi, R.; Zhu, Y.; Wang, Z.; Pan, H.; Wang, D. Machine learning algorithm to predict mortality in critically ill patients with sepsis-associated acute kidney injury. *Scientific Reports* **2023**, *13*, 5223.
75. Tonekaboni, S.; Joshi, S.; McCradden, M.D.; Goldenberg, A. What clinicians want: Contextualizing explainable machine learning for clinical end use. *Proceedings of the 4th Machine Learning for Healthcare Conference* **2019**, *106*, 359–380.
76. Bento, V.; Kohler, M.; Diaz, P.; Mendoza, L.; Pacheco, M.A. Improving deep learning performance by using Explainable Artificial Intelligence (XAI) approaches. *Discover Artificial Intelligence* **2021**, *1*, 1–11.
77. Ou, S.M.; Tsai, M.T.; Lee, K.H.; Tseng, W.C.; Yang, C.Y.; Chen, T.H.; Bin, P.J.; Chen, T.J.; Lin, Y.P.; Sheu, W.H.H.; others. Prediction of the risk of developing end-stage renal diseases in newly diagnosed type 2 diabetes mellitus using artificial intelligence algorithms. *BioData Mining* **2023**, *16*, 8.

 doi:10.20944/preprints202408.1702.v1

78. Ahmad, M.; Eckert, C.; Teredesai, A.; Raymer, M.; Ramaswamy, S. Interpretable machine learning models for prediction of adverse cardiac events. *Nature Communications* **2018**, *9*, 4246.
79. Ravindran, K.; Jothimurugesan, D. A survey of explainable artificial intelligence in healthcare. *Artificial Intelligence in Medicine* **2023**, *107*, 102403.
80. Mienye, I.D.; Ainah, P.K.; Emmanuel, I.D.; Esenogho, E. Sparse noise minimization in image classification using Genetic Algorithm and DenseNet. 2021 Conference on Information Communications Technology and Society (ICTAS). IEEE, 2021, pp. 103–108.
81. McKinney, S.M.; Sieniek, M.; Godbole, V.; Godwin, J.; Antropova, N.; Ashrafian, H.; et al.. International evaluation of an AI system for breast cancer screening. *Nature* **2020**, *577*, 89–94.
82. DeGrave, A.J.; Janizek, J.D.; Lee, S.I. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nature Machine Intelligence* **2021**, *3*, 610–619.
83. Baumgartner, C.F.; Biffi, C.; Bakas, S.; et al.. Visual feature attribution using Wasserstein GANs. *Medical Image Analysis* **2018**, *46*, 60–73.
84. Gargeya, R.; Leng, T. Automated identification of diabetic retinopathy using deep learning. *Ophthalmology* **2017**, *124*, 962–969.
85. Albahri, A.S.; Duhaim, A.M.; Fadhel, M.A.; Alnoor, A.; Baqer, N.S.; Alzubaidi, L.; Albahri, O.S.; Alamoodi, A.H.; Bai, J.; Salhi, A.; others. A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion. *Information Fusion* **2023**, *96*, 156–191.
86. Hauser, K.; Kurz, A.; Haggenmüller, S.; Maron, R.C.; von Kalle, C.; Utikal, J.S.; Meier, F.; Hobelsberger, S.; Gellrich, F.F.; Sergon, M.; others. Explainable artificial intelligence in skin cancer recognition: A systematic review. *European Journal of Cancer* **2022**, *167*, 54–69.
87. Tschandl, P.; Rosendahl, C.; Kittler, H. Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: an open, web-based, international, diagnostic study. *The Lancet Oncology* **2019**, *20*, 938–947.
88. Arcadu, F.; Benmansour, F.; Maunz, A.; Willis, J.; Haskova, Z.; Prunotto, M. Deep learning algorithm predicts diabetic retinopathy progression in individual patients. *NPJ digital medicine* **2019**, *2*, 92.
89. Ribeiro, M.M.; Rocha, L.B.; Leal, C.C.; Santos, A.R.; Pires, D.S.; Santana, E.; Züllig, J.S. Automatic detection of arrhythmias from imbalanced data using machine learning techniques. *Expert Systems with Applications* **2020**, *158*, 113551.
90. Goodman, B.; Flaxman, S. European Union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine* **2017**, *38*, 50–57.
91. Mienye, I.D.; Sun, Y. Effective feature selection for improved prediction of heart disease. Pan-African Artificial Intelligence and Smart Systems Conference. Springer, 2021, pp. 94–107.
92. Xu, T.; White, J.; Kalkan, S.; Gunes, H. Investigating bias and fairness in facial expression recognition. Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16. Springer, 2020, pp. 506–523.
93. Wang, Y.; Ma, W.; Zhang, M.; Liu, Y.; Ma, S. A Survey on the Fairness of Recommender Systems. *ACM Transactions on Information Systems* **2023**, *41*, 1–43. doi:10.1145/3547333.
94. Mary, J.; Calauzenes, C.; El Karoui, N. Fairness-aware learning for continuous attributes and treatments. International Conference on Machine Learning. PMLR, 2019, pp. 4382–4391.
95. Harris, C.G. Combining Human-in-the-Loop Systems and AI Fairness Toolkits to Reduce Age Bias in AI Job Hiring Algorithms. 2024 IEEE International Conference on Big Data and Smart Computing (BigComp). IEEE, 2024, pp. 60–66.
96. Char, D.S.; Shah, N.H.; Magnus, D. Implementing machine learning in health care—addressing ethical challenges. *New England Journal of Medicine* **2018**, *378*, 981–983.
97. Jobin, A.; Ienca, M.; Vayena, E. The global landscape of AI ethics guidelines. *Nature Machine Intelligence* **2019**, *1*, 389–399.
98. Garrido-Muñoz, I.; Montejo-Ráez, A.; Martínez-Santiago, F.; Ureña-López, L.A. A Survey on Bias in Deep NLP. *Applied Sciences* **2021**, *11*. doi:10.3390/app11073184.
99. Rajkomar, A.; Hardt, M.; Howell, M.D.; Corrado, G.; Chin, M.H. Ensuring fairness in machine learning to advance health equity. *Annals of internal medicine* **2018**, *169*, 866–872.
100. Mitchell, S.; Potash, E.; Barocas, S.; D'Amour, A.; Lum, K. Algorithmic fairness: Choices, assumptions, and definitions. *Annual review of statistics and its application* **2021**, *8*, 141–163.

 doi:10.20944/preprints202408.1702.v1

101. He, J.; Hao, Y.; Wang, X. An interpretable aid decision-making model for flag state control ship detention based on SMOTE and XGBoost. *Journal of Marine Science and Engineering* **2021**, *9*, 156.
102. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence* **2019**, *1*, 206–215.
103. Dewi, C.; Tsai, B.J.; Chen, R.C. Shapley additive explanations for text classification and sentiment analysis of internet movie database. Asian Conference on Intelligent Information and Database Systems. Springer, 2022, pp. 69–80.
104. Daries, J.P.; Reich, J.; Waldo, J.; et al.. Privacy, anonymity, and big data in the social sciences. *Communications of the ACM* **2014**, *57*, 56–63.