Informatics in Medicine Unlocked 51 (2024) 101587

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: www.elsevier.com/locate/imu

A survey of explainable artificial intelligence in healthcare: Concepts, applications, and challenges

Ibomoiye Domor Mienye [a], George Obaido [b,*], Nobert Jere [c], Ebikella Mienye [a], Kehinde Aruleba [d], Ikiomoye Douglas Emmanuel [e], Blessing Ogbuokiri [f]

[a] Institute for Intelligent Systems, University of Johannesburg, Auckland Park, Johannesburg, 2006, South Africa
[b] Center for Human-Compatible Artificial Intelligence (CHAI), Berkeley Institute for Data Science (BIDS), University of California, Berkeley, CA, 94720, United States
[c] Department of Computer Science, University of Fort Hare, Alice Campus, Alice, Eastern Cape, South Africa
[d] School of Computing and Mathematical Sciences, University of Leicester, Leicester, LE1 7RH, United Kingdom
[e] School of Science, Engineering and Environment, University of Salford, Salford, United Kingdom
[f] Department of Computer Science, Brock University, Niagara Region, St. Catharines, ON, L2S 3A1, Canada

ARTICLE INFO

Keywords:
AI
Bias
Ethics
Fairness
Healthcare
Machine learning

ABSTRACT

Explainable AI (XAI) has the potential to transform healthcare by making AI-driven medical decisions more transparent, reliable, and ethically compliant. Despite its promise, the healthcare sector faces several challenges, including the need to balance 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, offering valuable insights for researchers and practitioners. The insights and recommendations presented aim to guide future research and policy-making, fostering the development of transparent, trustworthy, and effective AI-driven solutions.

Contents

1. Introduction 2
2. Related works 2
3. Overview of explainable AI 3
   3.1. Transparency 3
   3.2. Trustworthiness 3
   3.3. Accountability 3
4. XAI techniques and methods 3
   4.1. Model-specific techniques 3
      4.1.1. Decision trees and rule-based systems 4
      4.1.2. Attention mechanisms in neural networks 4
      4.1.3. Convolutional Neural Networks 4
      4.1.4. Bayesian networks 4
   4.2. Model-agnostic techniques 5
      4.2.1. SHapley Additive exPlanations 5
      4.2.2. Local Interpretable Model-agnostic Explanations 6
      4.2.3. Partial Dependence Plots 6
      4.2.4. Individual Conditional Expectation plots 6
      4.2.5. Surrogate models 6
      4.2.6. Counterfactual explanations 7

* Corresponding author.
E-mail addresses: ibomoiyem@uj.ac.za (I.D. Mienye), gobaido@berkeley.edu (G. Obaido), njere@ufh.ac.za (N. Jere), 219105099@student.uj.ac.za (E. Mienye), ka388@leicester.ac.uk (K. Aruleba), i.d.emmanuel@edu.salford.ac.uk (I.D. Emmanuel), bogbuokiri@brocku.ca (B. Ogbuokiri).

https://doi.org/10.1016/j.imu.2024.101587
Received 11 August 2024; Received in revised form 3 October 2024; Accepted 10 October 2024
Available online 16 October 2024

4.2.7. Permutation feature importance ... 7
5. Evaluation metrics for XAI ... 7
5.1. Fidelity ... 7
5.2. Interpretability ... 7
5.3. Stability ... 7
5.4. Completeness ... 7
5.5. Sparsity ... 8
6. Responsible AI in healthcare ... 8
6.1. Ethical considerations ... 8
6.2. Accountability and transparency ... 8
6.3. Fairness and bias mitigation ... 8
6.4. Human-in-the-loop approaches ... 8
7. Applications of explainable AI in healthcare ... 9
7.1. Diagnostic tools and Clinical Decision Support Systems ... 9
7.2. Personalized medicine ... 9
7.3. Medical imaging ... 9
7.4. Remote diagnostics and telemedicine ... 10
8. Challenges and opportunities in implementing XAI in healthcare ... 10
8.1. Integration into clinical workflows ... 10
8.2. Regulatory compliance ... 10
8.3. Bias and fairness ... 11
8.4. Interpretability vs. Accuracy ... 11
8.5. Long-term impact on patient outcomes ... 11
9. Discussion and future research directions ... 11
10. Conclusion ... 12
CRediT authorship contribution statement ... 12
Declaration of competing interest ... 12
Acknowledgments ... 12
References ... 12

1. Introduction

Artificial Intelligence (AI) has significantly reshaped numerous sectors, with healthcare undergoing remarkable transformations [1,2]. The integration of AI technologies into healthcare systems has enabled advancements once considered unattainable. These advancements include enhanced diagnostic accuracy, personalized treatment plans, and improved operational efficiencies [3,4]. Machine learning (ML) and deep learning (DL), as subfields of AI, facilitate the processing and analysis of vast datasets, which are critical for predictive analytics and decision-making in clinical environments [5,6].

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 particularly important in healthcare, where understanding the rationale behind AI predictions is crucial for clinicians and patients [7–9]. 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, trust, and reliability of AI technologies in medical practice.

In the past, researchers have conducted various reviews on AI in healthcare. Some reviews have focused on the general application of AI models in healthcare [10–13], while others have addressed the potential of AI for predictive analytics in clinical settings [14,15] and drug discovery [2,16,17]. Additionally, some reviews include discussions on the ethical implications and regulatory challenges associated with AI in healthcare [3,18]. However, most reviews either focus on AI broadly or do not provide an in-depth analysis of XAI tailored to healthcare. Furthermore, 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 due to 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, while Section 3 discusses the fundamental concepts of XAI. Section 4 outlines various XAI methods and techniques, and Section 5 presents evaluation metrics used in XAI. Section 6 examines responsible AI in healthcare, and Section 7 explores the applications of XAI in healthcare. Section 8 identifies the challenges and opportunities in implementing XAI within clinical settings, and Section 9 provides a comprehensive discussion on future research directions, identifying areas of potential growth and innovation. Finally, Section 10 concludes the paper with a summary of key findings.

2. Related works

The application of AI in healthcare has seen rapid advancements in recent years, with studies focusing on its various implementations and impacts. For instance, Maleki and Forouzanfar [14] provided a detailed analysis of AI's capabilities in clinical settings, emphasizing its contributions to diagnostic accuracy and patient care improvements. Additionally, more recent work by Kalra et al. [19] examined the growing role of AI in medical diagnosis, particularly its integration with electronic health systems, and discussed the complexities of embedding AI into clinical workflows. Similarly, Liu et al. [20] presented an overview of AI applications in medicine, detailing various AI models and their effectiveness in disease prediction and management.

Meanwhile, XAI is a crucial subfield within AI that addresses the need for transparency and interpretability in AI models. Arrieta et al. [21] provided an extensive review of XAI techniques, categorizing them into model-specific and model-agnostic approaches, and highlighted the significant role XAI plays in fostering trust and accountability in AI-driven systems. Notable studies in XAI include Tosun et al. [22], which discussed computational approaches to XAI,

offering cutting-edge techniques for explainability. Similarly, Longo et al. [23] provided a forward-looking perspective on XAI 2.0, outlining new interdisciplinary research directions and addressing open challenges in AI model explainability.

Furthermore, Adadi and Berrada [24] explored the potential of XAI across multiple sectors, including healthcare, identifying core challenges in rendering AI models interpretable and proposing possible solutions. The role of Responsible AI, which encompasses ethical considerations like fairness, accountability, and transparency, has gained prominence. Dignum [25] outlined principles of Responsible AI, emphasizing the importance of embedding ethical standards into AI development processes. Additionally, Mienye et al. [26] discussed fairness in AI, including strategies for detecting and mitigating biases in healthcare ML models.

Recent advancements have also focused on model-specific XAI techniques, which provide explanations that are deeply integrated into the specific types of models being used. One such example is the work by Konstantinov and Utkin [27], which introduced new methods to improve the interpretability of gradient-boosting machines by employing parallel gradient boosting models. Their approach uses linear combinations of boosting models and includes Lasso-based techniques to update model weights, making it highly effective for diagnostic tools, particularly in areas like oncology and cardiology. Additionally, Raghavan [28] explored the application of XAI in deep learning models designed for medical imaging, where model-specific techniques like Grad-CAM and attention mechanisms provided real-time visual explanations for MRI and CT scan diagnostics.

Another recent development in XAI is the introduction of hybrid XAI techniques, which combine the strengths of both model-specific and model-agnostic approaches. This combination enhances the flexibility and scalability of AI models across different healthcare domains. For example, Khan et al. [29] demonstrated how hybrid XAI methods could be applied to both structured and unstructured medical data, improving explainability within clinical decision support systems. These hybrid approaches represent a new trend in XAI, addressing the limitations of purely model-agnostic or model-specific methods by providing both global and local explanations, thus enhancing the transparency of AI systems used in complex medical environments.

Meanwhile, Holzinger et al. [30] discussed the importance of human-in-the-loop (HITL) approaches in AI, particularly in healthcare. They argued that AI systems must not only be accurate but also provide explanations that are comprehensible to users to be trusted and widely adopted in clinical practice. Their work underscores the necessity of theoretical foundations for XAI and the practical role of HITL methods in ensuring effective AI integration.

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, Singla [31] reviewed the application of AI in healthcare, providing insights into the potential of AI but not addressing the interpretability challenges faced by healthcare professionals. Similarly, Esteva et al. [32] and Kaul et al. [33] focused on deep learning in healthcare, offering insights into its potential but not addressing the interpretability challenges associated with these models. 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 [34]. This is particularly important in fields such as healthcare, where understanding the reasoning behind AI decisions can directly impact patient outcomes and foster trust in the technology. The main goals of XAI include:

3.1. Transparency

Transparency involves making the AI decision-making process clear and understandable [35]. 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 building trust in AI-driven healthcare systems. 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 [36].

3.2. Trustworthiness

Trustworthiness involves building user trust through understandable and verifiable AI decisions [21]. Trust is essential for the widespread adoption of AI technologies, particularly in critical fields like healthcare [37]. 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 benefit from expert knowledge and feedback, leading to more accurate and reliable outcomes [38]. Additionally, frameworks that track and log AI decisions, providing an audit trail, help users review and understand the rationale behind AI-driven decisions, further enhancing trust.

3.3. 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 [37]. 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 [39]. Furthermore, the key terms and concepts associated with explainable AI are tabulated in Table 1.

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.

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:

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 [26].
Responsible AI	AI that takes into account societal values, morals, and ethical considerations, focusing on accountability, responsibility, and transparency [40].
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.
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 [37].
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 [41].
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 [42].
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 [43].
Ethical AI	The practice of designing and deploying AI systems in ways that are aligned with ethical principles, such as fairness, accountability, and transparency.
Data Privacy	The protection of personal data used in AI systems, ensuring that sensitive information is handled securely and ethically.

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 [44]. 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) [26]. 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 [26]. The structure of a typical decision tree is shown in Fig. 1.

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 [46–48]. Attention mechanisms assign different weights to different parts of the input, highlighting their relative importance in the final decision [49]. 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 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 [50]. 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 [51,52]. 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 [51].

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) [53]. 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.

Furthermore, model-specific techniques offer significant adaptability, especially in healthcare, where the interpretability and transparency of AI systems are critical. These methods are inherently designed to align with specific model architectures, which allows for deeper insights into how these models make predictions. Such model-specific techniques exploit the structural and functional characteristics of individual models to provide more granular explanations. This adaptability makes them especially suitable for healthcare applications where understanding the model's decision-making process is essential for clinicians. For example, decision trees and rule-based systems are

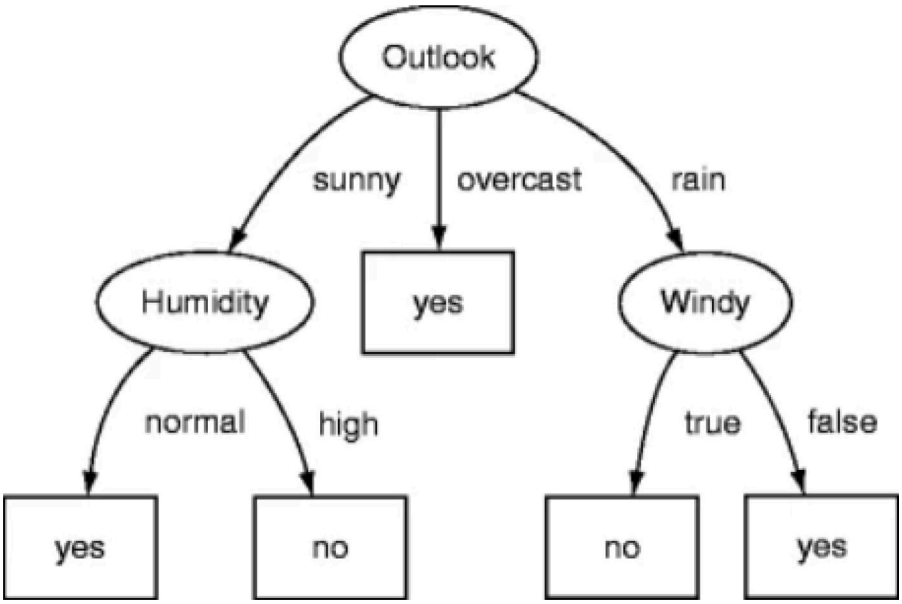

Fig. 1. Example of a decision tree [45].

not only inherently interpretable but also adaptable to various types of clinical data, including both structured data (such as lab results) and unstructured data (such as clinical notes) [54,55]. The clear, step-by-step nature of decision trees, which follow a flow of if-then rules, is particularly useful for diagnostic decision-making. These models can be easily adapted to new data sources and clinical environments without losing their interpretability.

Similarly, attention mechanisms in neural networks offer a high level of adaptability for interpreting complex and high-dimensional data, such as medical images or genomic data [28,56]. The ability of attention mechanisms to highlight the most relevant parts of the input data makes them highly effective in medical applications like MRI scans or pathology slides, where understanding the critical areas influencing a prediction is vital. Furthermore, attention mechanisms can be adapted to different neural network architectures, making them versatile across various healthcare domains, from radiology to genomics. CNNs, particularly when enhanced with techniques like Grad-CAM, provide interpretable visual outputs that highlight the regions in an image most responsible for the model's prediction [28]. This adaptability makes CNNs and their interpretability techniques suitable for a wide range of healthcare applications, including radiology, dermatology, and ophthalmology. The use of visual explanations is crucial for clinical professionals, as it allows them to cross-verify AI-generated outputs with their clinical expertise, thereby improving trust in AI systems.

Additionally, Bayesian networks are highly adaptable model-specific techniques and are valuable for modeling uncertainty and understanding probabilistic relationships between variables [53]. In healthcare, where uncertainty is often present in patient outcomes or disease progression, Bayesian networks provide a clear and interpretable way to model these uncertainties. Their flexibility in incorporating both expert knowledge and data-driven insights makes them particularly useful for developing clinical decision support systems that need to account for variable and uncertain clinical environments.

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 [54]. 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 [54]. 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 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 [57]. 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 [58]. An example architecture for integrating SHAP with an ML model is shown in Fig. 2.

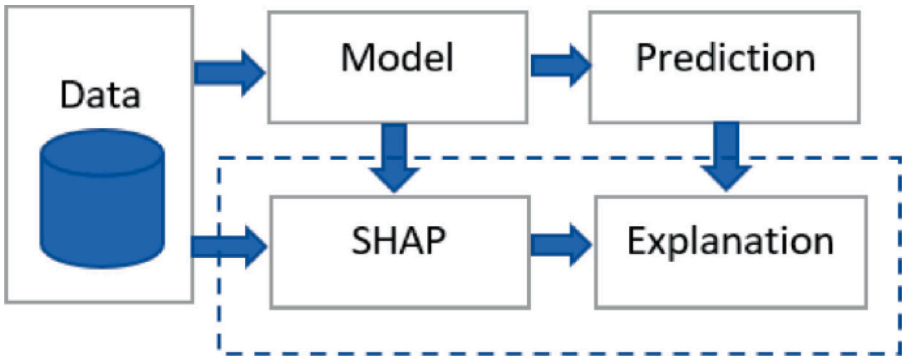

Fig. 2. SHAP-ML model architecture [59].

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 [60]. 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 [60,61]. Algorithm 1 summarizes the LIME process:

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 [62]. 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 [63].

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 [64]. For an instance i, the ICE curve for a feature x_j is given by:

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

ICE plots provide a more granular view of feature effects, revealing heterogeneity in the model's behavior across different instances [65]. 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.

4.2.5. Surrogate models

Surrogate models are interpretable models that approximate the predictions of a more complex, black-box model [66]. These models are typically used to explain the behavior of machine learning models that are inherently difficult to interpret, such as deep learning and ensemble models. The primary purpose of a surrogate model is to maintain interpretability while mimicking the performance of the black-box model as closely as possible. Given a black-box model f and a dataset X, a surrogate model g is trained to approximate the predictions of f for all instances $x \in X$:

$$g(x) \approx f(x) \quad \forall x \in X. \tag{12}$$

To create a surrogate model, an original complex model f is first trained on the dataset. Then, the surrogate model is trained on the predictions of f, using the same input features but with a simpler, more interpretable structure such as a decision tree or linear regression model [67]. The surrogate model g is interpretable because its structure, such as the rules in a decision tree, can be easily visualized and

understood by humans. This method is widely used to explain complex models in domains such as healthcare and finance. The fidelity of the surrogate model is critical, as it determines how well the simplified model approximates the behavior of the black-box model. Mathematically, the fidelity is often measured by comparing the predictions of g with those of f using metrics such as mean squared error (MSE):

$$\text{Fidelity} = \frac{1}{n}\sum_{i=1}^{n}(g(x_i) - f(x_i))^2. \tag{13}$$

Higher fidelity indicates a closer approximation to the black-box model.

4.2.6. Counterfactual explanations

Counterfactual explanations provide insights into what changes to the input data would lead to a different outcome from a machine learning model [68]. This technique is useful for identifying actionable changes in real-world applications such as loan approval, healthcare, and legal decisions. A counterfactual explanation answers the question: "What minimal change in the input would have resulted in a different prediction?" [69]. Given a black-box model f, an instance x, and a desired output y_{desired}, the goal of counterfactual reasoning is to find a modified instance x' such that:

$$f(x') = y_{\text{desired}} \quad \text{and} \quad \text{distance}(x, x') \quad \text{is minimized.} \tag{14}$$

The distance metric, such as the Euclidean distance or Manhattan distance, ensures that x' remains similar to x, preserving the interpretability and practicality of the explanation. For instance, in a healthcare setting, if a model predicts that a patient has a high risk of developing heart disease, a counterfactual explanation might suggest that lowering the patient's cholesterol level by a specific amount could reduce their risk. This actionable insight helps users understand how they can change an outcome by modifying specific features, making the model's decision-making process more transparent.

4.2.7. Permutation feature importance

Permutation feature importance is a model-agnostic technique used to evaluate the importance of individual features in a ML model [62]. This method works by permuting the values of each feature in the dataset and measuring the resulting decrease in model performance. Features that are important for the model's predictions will cause a large drop in performance when permuted, while less important features will have little to no effect [70]. Let f be the trained model, and X be the dataset with x_j as the j-th feature. The model's baseline performance is denoted by $\text{Perf}(f, X)$. The feature importance of x_j is calculated by permuting its values x_j^{perm} and computing the decrease in performance:

$$\text{Importance}(x_j) = \text{Perf}(f, X) - \text{Perf}(f, X^{\text{perm}(x_j)}). \tag{15}$$

In healthcare, this technique can be applied to assess the importance of clinical features such as blood pressure, cholesterol levels, and age in predicting the risk of heart disease [71]. Clinicians can gain valuable insights into the underlying factors driving the model's predictions by identifying the most influential features, which can be used to inform treatment decisions.

5. Evaluation metrics for XAI

Evaluating the effectiveness of XAI techniques is crucial to ensure that the explanations provided are useful, accurate, and actionable in healthcare. XAI evaluation requires metrics that assess the performance of the AI models and the quality of the interpretability and explanations. In this section, key evaluation metrics for XAI are discussed, including fidelity, interpretability, stability, and completeness, each of which helps determine the practical utility of XAI methods in healthcare applications.

5.1. Fidelity

Fidelity measures how well the explanation aligns with the predictions of the original black-box model [72]. A high-fidelity explanation closely approximates the behavior of the underlying model, ensuring that the explanation accurately reflects the true decision-making process. Mathematically, fidelity can be expressed as the degree to which the surrogate model g mimics the original model f. Fidelity F is defined as:

$$F = \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(g(x_i) = f(x_i)), \tag{16}$$

where $g(x_i)$ is the prediction of the surrogate interpretable model for instance x_i, and $f(x_i)$ is the prediction of the original model. $\mathbb{1}$ is an indicator function that equals 1 if the predictions match and 0 otherwise, and n represents the number of data instances. High fidelity is critical for the trustworthiness of the explanations [73].

5.2. Interpretability

Interpretability measures the ease with which human users can understand the model's decision-making process. This metric is subjective and often involves user studies, where clinicians or healthcare professionals assess the clarity of explanations [6]. While no single mathematical formulation exists for interpretability, proxy measures such as the complexity of the explanation are often used. For example, the complexity of a decision tree C can be represented as:

$$C = \sum_{i=1}^{L}\text{Depth}(n_i), \tag{17}$$

where L is the number of leaf nodes in the tree and $\text{Depth}(n_i)$ is the depth of node n_i. Lower complexity generally correlates with higher interpretability.

5.3. Stability

Stability refers to the consistency of explanations when there are small perturbations in the input data. If small changes in the input lead to drastically different explanations, the XAI method may be unreliable [74]. It is crucial in healthcare, where clinicians need to trust that similar patients will receive similar explanations. Stability can be measured by the variance in the explanations across neighboring instances. Assuming ϵ_i be the explanation for instance x_i and ϵ_j the explanation for its neighbor x_j, the stability S can be computed as:

$$S = \frac{1}{n}\sum_{i=1}^{n}\|\epsilon_i - \epsilon_j\|_2, \tag{18}$$

where $\|\cdot\|_2$ represents the Euclidean distance between the explanations. Lower values of S indicate higher stability.

5.4. Completeness

Completeness measures how much of the original model's behavior is captured by the explanation [75]. This metric is important when using feature attribution methods like SHAP or LIME. Completeness can be defined as the proportion of variance in the model's predictions that is explained by the XAI method [76]. Let y_i be the model's prediction for instance x_i and $\hat{y}_i$ be the prediction of the interpretable explanation model. The completeness C is calculated as:

$$C = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{19}$$

where $\bar{y}$ is the mean prediction of the model. A completeness score close to 1 indicates that the XAI method captures most of the original model's behavior.

5.5. Sparsity

Sparsity measures the number of features used in an explanation. Sparse explanations are generally more interpretable because they focus on a small set of key features rather than overwhelming the user with many variables [77]. Sparsity S_p can be quantified by counting the number of non-zero feature attributions in a model like SHAP:

$$S_p = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(\phi_j^{(i)} \neq 0), \tag{20}$$

where $\phi_j^{(i)}$ represents the Shapley value for feature j in instance i, and $\mathbb{1}$ is the indicator function that counts non-zero values. Lower sparsity indicates a more focused, interpretable explanation.

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. These systems must be designed to uphold patient privacy, data security, and the integrity of clinical decision-making processes. Regulations such as the Health Insurance Portability and Accountability Act (HIPAA) in the United States and the General Data Protection Regulation (GDPR) in Europe set strict standards for the use of sensitive health information, ensuring its protection throughout the data processing lifecycle [40,78]. Robust data encryption, anonymization, and stringent access controls are necessary to comply with these regulations, preventing unauthorized access or data breaches.

Furthermore, in 2024, the European Union (EU) published an AI Act, the world's first comprehensive AI regulation [79]. With the advent of the EU AI Act, ethical AI deployment is subject to even more stringent oversight. This legislation categorizes AI systems based on their risk levels and mandates that high-risk AI systems, such as those used in healthcare, must meet strict criteria for transparency, accountability, and fairness [80]. Under this act, healthcare AI systems must provide clear documentation of their decision-making processes, allowing clinicians to audit AI-generated predictions and ensuring that patients can contest AI-driven decisions that affect their care. The EU AI Act also emphasizes the need for human oversight, ensuring that AI systems do not operate in a completely autonomous manner, thus safeguarding against potential ethical violations.

Ethical AI practices must also include obtaining informed consent from patients regarding the use of their data. This involves ensuring that patients understand how their data will be used, stored, and protected and offering them the option to opt-out if necessary. With the increased complexity of AI systems, explainability becomes critical for informed consent. By providing interpretable explanations of AI decisions, patients and clinicians can trust the AI's role in treatment pathways [23].

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 [81]. Transparency in AI systems involves making the decision-making processes visible and understandable to users [82]. This can be achieved through various XAI techniques, such as LIME and SHAP, which show how AI models arrive at their conclusions.

In addition to technical transparency, organizational accountability is gaining prominence. Healthcare institutions must ensure they implement governance frameworks that regularly audit and evaluate AI systems, holding developers and healthcare professionals accountable for AI-driven decisions [83]. Additionally, global frameworks such as the EU AI Act are pushing for stricter transparency and accountability measures for high-risk AI systems, including those in healthcare, by requiring that all decisions be explainable and auditable by design [80]. By aligning technical transparency with legal and organizational frameworks, AI systems can achieve a higher level of accountability, ultimately leading to more reliable and ethically aligned healthcare practices.

6.3. Fairness and bias mitigation

Fairness in AI refers to the impartiality and equity in AI decision-making processes [84]. 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 [85]. Techniques such as re-sampling the training data, adjusting model parameters, and incorporating fairness constraints during model training are essential to mitigate these biases [43]. 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.

Recent research has shown that intersectional fairness is essential to address biases arising from overlapping social categories, such as race, gender, and socioeconomic status [86]. Approaches such as counterfactual fairness are emerging as vital tools to ensure that AI decisions remain fair under different hypothetical scenarios [87]. By continuously monitoring models with techniques like algorithmic impact assessments, healthcare institutions can ensure long-term fairness, maintaining the balance between performance and equity in patient outcomes. Furthermore, fairness is not just about technical accuracy but about ensuring equity in access to healthcare resources, treatments, and medical interventions, thus enabling inclusivity in healthcare systems.

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 [42]. 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 [88].

In recent studies, the adaptive HITL frameworks have gained traction, where the system continuously learns from human feedback, adapting its predictions to align more closely with clinical expertise [42,89,90]. Meanwhile, HITL is evolving to involve patient-in-the-loop (PITL) approaches, especially in patient-centered care, where patients' preferences and values are incorporated into the decision-making loop, improving shared decision-making processes [91]. By adopting such inclusive and dynamic frameworks, HITL approaches will not only boost trustworthiness but also facilitate the personalized care that modern healthcare systems aspire to provide.

7. 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.

7.1. 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, thereby improving the accuracy and transparency of diagnostics [92–94]. Recent studies, such as that by Lundberg and Lee [95], 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.

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. [96] used LIME to interpret model predictions, enhancing the transparency and trustworthiness of the CDSS. Similarly, Bedoya et al. [15] employed SHAP values in CDSS for predicting hospital readmissions, helping healthcare providers tailor care plans to individual patients. Another study by Tonekaboni et al. [97] 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. [98] 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.

7.2. 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 [99].

Additionally, XAI has shown potential in managing cardiovascular diseases. A study by Alkhamis et al. [100] 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 Thiruvenkadam et al. [101] used XAI to interpret deep learning models used for predicting brain tumors based on MRI, enabling clinicians to understand and validate the AI's recommendations.

Mienye and Jere [54] have utilized the SHAP technique in predicting outcomes in personalized treatments, particularly in cases involving chronic diseases. They demonstrated how SHAP values can assist in understanding the contribution of various factors, such as patient lifestyle and medical results to treatment outcomes. This transparency in treatment recommendations improves both patient and clinician trust in the AI system. Similarly, Sirapangi and Gopikrishnan [102] developed a multimodal personalized treatment model that uses feature selection and XAI to explain why specific treatments are recommended based on real-time patient data. This work highlights how the integration of XAI in personalized medicine improves treatment accuracy and aids in better patient-clinician communication, enhancing overall healthcare quality.

Furthermore, Khater et al. [103] explored how XAI techniques can be used to personalize lifestyle recommendations in healthcare. They used SHAP to interpret the contribution of factors in obesity detection. This research illustrates the potential of XAI to provide personalized preventive healthcare advice in addition to treatment recommendations, thereby expanding its application beyond traditional medicinal approaches.

7.3. 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 [104,105].

Recent studies have highlighted the effectiveness of XAI in various medical imaging applications. For instance, Ahmed et al. [106] 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. [107] 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. [108] 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 [109] showing how heatmaps can help ophthalmologists understand and trust AI-driven diagnoses.

Raghavan [28] demonstrated the application of Grad-CAM in diagnosing early-stage breast cancer from mammograms, showing how attention-based techniques can provide visual evidence that supports the AI's prediction. This allows radiologists to cross-check AI predictions with their expertise, reducing diagnostic errors and improving confidence in AI-assisted medical imaging. Their study also emphasized the importance of XAI in building trust between AI systems and medical professionals, especially when dealing with life-critical decisions.

Additionally, Grad-CAM has been employed in MRI-based brain tumor detection by Mahesh et al. [110], who demonstrated the use of these techniques to visually explain which brain regions are influencing the model's decision. This not only improves diagnostic accuracy but also allows healthcare professionals to better understand AI predictions. Similarly, Wang et al. [111] utilized SHAP values to explain predictions in diabetic retinopathy detection, offering an interpretable view of which retinal features are most relevant to the diagnosis.

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	[92]
	Cardiovascular diseases	Detecting patterns in ECGs indicative of heart diseases	[93]
	Pneumonia detection	Using SHAP values for pneumonia diagnosis from chest X-rays	[95]
	Diabetic retinopathy	Interpreting retinal images to predict diabetic retinopathy	[117]
	Neurological disorders	Diagnosing Alzheimer's using attention mechanisms on MRI scans	[94]
	Arrhythmia detection	Enhancing transparency in deep learning models for arrhythmia diagnosis	[118]
	Sepsis prediction	Using LIME to explain sepsis prediction in ICU patients	[96]
	Hospital readmission prediction	Employing SHAP values to predict hospital readmissions	[15]
Personalized medicine	Heart disease	Personalized treatment recommendations based on electronic health records	[100]
	Oncology	Tailoring chemotherapy treatments using genetic profiles and clinical data	[99]
	Cardiovascular diseases	Predicting risk of cardiac events and tailoring prevention strategies	[101]
	Gene therapies	Recommending gene editing techniques based on genomic data	[119]
	Chronic disease management	Using SHAP for personalized chronic disease treatment recommendations	[54]
	Multimodal treatment model	Explaining treatment recommendations with real-time data using XAI	[102]
	Obesity detection	Utilizing SHAP for lifestyle recommendation in obesity detection	[103]
Medical imaging	Breast cancer	Applying attention maps in deep learning models for mammogram analysis	[106]
	COVID-19 detection	Using saliency maps to diagnose COVID-19 from chest X-rays	[107]
	Brain tumors	Visualizing brain regions in MRI scans with Grad-CAM for tumor classification	[108]
	Diabetic retinopathy	Highlighting retinal areas in AI predictions for diabetic retinopathy diagnosis	[109]
	Breast cancer	Using Grad-CAM for early-stage breast cancer diagnosis in mammograms	[28]
	Brain tumors	Grad-CAM used for MRI-based brain tumor detection	[110]
	Diabetic retinopathy	SHAP used for explaining diabetic retinopathy predictions	[111]
Remote diagnostics and telemedicine	Respiratory diseases	Integrating XAI in telemedicine platforms for respiratory disease diagnosis	[120]
	Dermatology	Employing XAI for diagnosing skin conditions via mobile devices	[121]
	Ophthalmology	Interpreting AI models in telemedicine for eye disease diagnosis from retinal images	[114]
	Real-time diagnostics	XAI tools in real-time diagnostic tools for remote regions	[115]
	Respiratory diseases	Using SHAP for COVID-19 respiratory disease diagnosis in telemedicine	[116]

7.4. 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 [112]. For example, in dermatology, Ramoliya et al. [113] 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. Similarly, XAI enhances the trustworthiness of telemedicine platforms for diagnosing eye diseases from retinal images, as demonstrated by Arcadu et al. [114], improving the quality of care in underserved regions.

In telemedicine, Patel et al. [115] demonstrated how XAI tools such as LIME are integrated into real-time diagnostic tools used by healthcare workers in remote regions. These tools provide interpretable diagnostics, such as explaining skin conditions or respiratory issues detected through mobile phones or wearable devices. This research highlighted the importance of clear, interpretable AI tools in empowering remote healthcare providers who may not have access to specialized diagnostic expertise.

Another important study by Awotunde et al. [116] emphasized the need for real-time XAI in telemedicine platforms, particularly during the COVID-19 pandemic. The study showed how SHAP values were used to explain AI-driven diagnoses of respiratory diseases, providing clarity on which lung features were most indicative of infection in chest X-rays. The use of XAI in telemedicine is not only vital for improving diagnostic accuracy but also for building trust between patients, providers, and AI tools in geographically isolated areas. Table 2 summarizes the different XAI applications in healthcare.

8. 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:

8.1. 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 [122]. 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.

8.2. 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 [78]. 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 [123].

8.3. 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 [124]. 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 [125,126].

8.4. 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 [127]. 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 [128]. 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 [37,129].

8.5. 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.

9. Discussion and future research directions

This study has identified that one of the main benefits of XAI in healthcare is enhancing 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 [59]. 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 [128,129]. Additionally, 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 [78].

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 [43,125].

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

- **Developing inherently interpretable models**: Future research should prioritize the development of models that are transparent by design, reducing reliance on post-hoc explanation techniques like SHAP and LIME. This shift towards inherently interpretable models, such as generalized additive models (GAMs), rule-based learning models, and attention-based architectures, can provide direct insights into decision-making processes [130]. Inherently interpretable models are more trustworthy and better suited for real-time clinical decision-making, where transparency and interpretability are essential.
- **Integrating causal inference techniques**: Moving beyond correlation-based methods like SHAP and LIME, future research should emphasize the integration of causal inference techniques. Causal XAI can uncover the cause-and-effect relationships between features and outcomes, providing clinicians with more reliable explanations [131]. This approach is valuable in healthcare settings, where understanding the underlying causal factors can enhance clinical decisions and lead to better patient outcomes.
- **Advancing visualization tools and hybrid models**: Future work should focus on improving the visual representations of AI explanations. Hybrid models that combine interpretable components with complex, high-performance models could leverage advanced visualization techniques such as interactive dashboards and 3D visualizations. These tools can make AI-driven insights more comprehensible for healthcare professionals, facilitating their integration into clinical workflows.
- **Exploring real-time interpretability and natural language processing**: Real-time interpretability in clinical settings is an emerging area that requires further attention. Research should explore techniques that allow AI systems to provide instant, context-aware explanations during patient interactions. Moreover, the use of NLP for generating human-readable explanations could improve the accessibility and usability of XAI in healthcare, allowing clinicians to better understand AI-generated outputs in a natural language format.

- **Enhancing model robustness and generalization ability**: Robustness and generalization remain key challenges in deploying XAI techniques across diverse clinical environments. Future research should explore how XAI models can handle heterogeneous healthcare data while maintaining consistency in explanations across varied clinical settings. This will ensure that AI systems remain reliable and interpretable in real-world, dynamic healthcare environments.
- **Investigating ethical considerations and regulatory compliance**: Future studies should investigate the intersection of ethical concerns, such as bias mitigation and patient privacy, with the evolving regulatory landscape. With the introduction of frameworks like the EU AI Act, research must explore how XAI models can meet compliance requirements while maintaining their interpretability and effectiveness.

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.

10. Conclusion

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 analyzed, offering substantial contributions to the existing literature on healthcare AI. 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, with the ultimate goal of creating more transparent, trustworthy, and effective AI-driven healthcare solutions.

CRediT authorship contribution statement

Ibomoiye Domor Mienye: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **George Obaido:** Writing – review & editing, Writing – original draft, Visualization, Project administration, Conceptualization. **Nobert Jere:** Writing – review & editing, Writing – original draft, Visualization, Supervision. **Ebikella Mienye:** Writing – review & editing, Writing – original draft, Supervision. **Kehinde Aruleba:** Writing – review & editing, Writing – original draft. **Ikiomoye Douglas Emmanuel:** Writing – review & editing, Writing – original draft, Validation. **Blessing Ogbuokiri:** Writing – review & editing, Writing – original draft.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research received no specific grants from funding agencies in the public, commercial, or not-for-profit sectors.

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. Adv Eng Inform 2023;55:101874. http://dx.doi.org/10.1016/j.aei.2023.101874.

[2] Obaido G, Mienye ID, Egbelowo OF, Emmanuel ID, Ogunleye A, Ogbuokiri B, et al. Supervised machine learning in drug discovery and development: Algorithms, applications, challenges, and prospects. Mach Learn Appl 2024;17:100576. http://dx.doi.org/10.1016/j.mlwa.2024.100576.

[3] Lesley U, Kuratomi Hernández A. Improving XAI explanations for clinical decision-making – Physicians' perspective on local explanations in healthcare. In: Lecture notes in computer science. Springer Nature Switzerland; 2024, p. 296–312. http://dx.doi.org/10.1007/978-3-031-66535-6_32.

[4] Khalifa M, Albadawy M. AI in diagnostic imaging: Revolutionising accuracy and efficiency. Comput Methods Programs Biomed Update 2024;5:100146. http://dx.doi.org/10.1016/j.cmpbup.2024.100146.

[5] Lu S-C, Swisher CL, Chung C, Jaffray D, Sidey-Gibbons C. On the importance of interpretable machine learning predictions to inform clinical decision making in oncology. Front Oncol 2023;13. http://dx.doi.org/10.3389/fonc.2023.1129380.

[6] Mienye ID, Jere N. Deep learning for credit card fraud detection: A review of algorithms, challenges, and solutions. IEEE Access 2024;12:96893–910. http://dx.doi.org/10.1109/ACCESS.2024.3426955.

[7] Hulsen T. Explainable artificial intelligence (XAI): Concepts and challenges in healthcare. AI 2023;4(3):652–66. http://dx.doi.org/10.3390/ai4030034, URL https://www.mdpi.com/2673-2688/4/3/34.

[8] Borys K, Schmitt YA, Nauta M, Seifert C, Krämer N, Friedrich CM, et al. Explainable AI in medical imaging: An overview for clinical practitioners – Beyond saliency-based XAI approaches. Eur J Radiol 2023;162:110786. http://dx.doi.org/10.1016/j.ejrad.2023.110786.

[9] Obaido G, Ogbuokiri B, Chukwu CW, Osaye FJ, Egbelowo OF, Uzochukwu MI, et al. An improved ensemble method for predicting hyperchloremia in adults with diabetic ketoacidosis. IEEE Access 2024;12:9536–49. http://dx.doi.org/10.1109/ACCESS.2024.3351188.

[10] Alowais SA, Alghamdi SS, Alsuhebany N, Alqahtani T, Alshaya AI, Almohareb SN, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med Educ 2023;23(1). http://dx.doi.org/10.1186/s12909-023-04698-z.

[11] Rong G, Mendez A, Bou Assi E, Zhao B, Sawan M. Artificial intelligence in healthcare: Review and prediction case studies. Engineering 2020;6(3):291–301. http://dx.doi.org/10.1016/j.eng.2019.08.015, URL https://www.sciencedirect.com/science/article/pii/S2095809919301535.

[12] Loh HW, Ooi CP, Seoni S, Barua PD, Molinari F, Acharya UR. Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022). Comput Methods Programs Biomed 2022;226:107161. http://dx.doi.org/10.1016/j.cmpb.2022.107161.

[13] Li RC, Asch SM, Shah NH. Developing a delivery science for artificial intelligence in healthcare. npj Digit Med 2020;3(1). http://dx.doi.org/10.1038/s41746-020-00318-y.

[14] Maleki Varnosfaderani S, Forouzanfar M. The role of AI in hospitals and clinics: Transforming healthcare in the 21st century. Bioengineering 2024;11(4). http://dx.doi.org/10.3390/bioengineering11040337.

[15] Bedoya JCL, Castro JLA. Explainability analysis in predictive models based on machine learning techniques on the risk of hospital readmissions. Health Technol 2023;14(1):93–108. http://dx.doi.org/10.1007/s12553-023-00794-8.

[16] Blanco-Gonzalez A, Cabezon A, Seco-Gonzalez A, Conde Torres D, Antelo Riveiro P, Pineiro A, et al. The role of AI in drug discovery: Challenges, opportunities, and strategies. Pharmaceuticals 2023;16(6). http://dx.doi.org/10.3390/ph16060891, URL https://www.mdpi.com/1424-8247/16/6/891.

[17] Walters WP, Barzilay R. Critical assessment of AI in drug discovery. Expert Opin Drug Discovery 2021;16(9):937–47. http://dx.doi.org/10.1080/17460441.2021.1915982.

[18] Rogers WA, Draper H, Carter SM. Evaluation of artificial intelligence clinical applications: Detailed case analyses show value of healthcare ethics approach in identifying patient care issues. Bioethics 2021;35(7):623–33. http://dx.doi.org/10.1111/bioe.12885.

[19] Kalra N, Verma P, Verma S. Advancements in AI based healthcare techniques with FOCUS ON diagnostic techniques. Comput Biol Med 2024;179:108917. http://dx.doi.org/10.1016/j.compbiomed.2024.108917.

[20] Liu C, Tan Z, He M. Overview of artificial intelligence in medicine. In: Artificial intelligence in medicine. Springer Nature Singapore; 2022, p. 23–34. http://dx.doi.org/10.1007/978-981-19-1223-8_2.

[21] Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, et al. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion 2020;58:82–115. http://dx.doi.org/10.1016/j.inffus.2019.12.012, URL https://www.sciencedirect.com/science/article/pii/S1566253519308103.

[22] Tosun AB, Pullara F, Becich MJ, Taylor DL, Chennubhotla SC, Fine JL. HistoMapr: An explainable AI (xAI) platform for computational pathology solutions. In: Lecture notes in computer science. Springer International Publishing; 2020, p. 204–27. http://dx.doi.org/10.1007/978-3-030-50402-113.

[23] Longo L, Brcic M, Cabitza F, Choi J, Confalonieri R, Ser JD, et al. Explainable artificial intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions. Inf Fusion 2024;106:102301. http://dx.doi.org/10.1016/j.inffus.2024.102301, URL https://www.sciencedirect.com/science/article/pii/S1566253524000794.

[24] Adadi A, Berrada M. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access 2018;6:52138–60. http://dx.doi.org/10.1109/ACCESS.2018.2870052.

[25] Dignum V. Responsible artificial intelligence. In: Artificial Intelligence: Foundations, Theory, and Algorithms. Springer International Publishing; 2019, http://dx.doi.org/10.1007/978-3-030-30371-6.

[26] Mienye ID, Obaido G, Emmanuel ID, Ajani AA. A survey of bias and fairness in healthcare AI. In: 2024 IEEE 12th international conference on healthcare informatics. 2024, p. 642–50. http://dx.doi.org/10.1109/ICHI61247.2024.00103.

[27] Konstantinov AV, Utkin LV. Interpretable machine learning with an ensemble of gradient boosting machines. Knowl-Based Syst 2021;222:106993. http://dx.doi.org/10.1016/j.knosys.2021.106993.

[28] Raghavan K. Attention guided grad-CAM: an improved explainable artificial intelligence model for infrared breast cancer detection. Multimedia Tools Appl 2024;83(19):57551–78.

[29] Khan N, Nauman M, Almadhor AS, Akhtar N, Alghuried A, Alhudhaif A. Guaranteeing correctness in black-box machine learning: A fusion of explainable AI and formal methods for healthcare decision-making. IEEE Access 2024;12:90299–316. http://dx.doi.org/10.1109/ACCESS.2024.3420415.

[30] Holzinger A, Biemann C, Pattichis CS, Kell DB. What do we need to build explainable AI systems for the medical domain?. 2017, arXiv:1712.09923. URL https://arxiv.org/abs/1712.09923.

[31] Singla S. AI and IoT in healthcare. In: Internet of things use cases for the healthcare industry. Springer International Publishing; 2020, p. 1–23. http://dx.doi.org/10.1007/978-3-030-37526-3_1.

[32] Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, et al. A guide to deep learning in healthcare. Nature Med 2019;25(1):24–9. http://dx.doi.org/10.1038/s41591-018-0316-z.

[33] Kaul D, Raju H, Tripathy BK. Deep learning in healthcare. In: Studies in big data. Springer International Publishing; 2021, p. 97–115. http://dx.doi.org/10.1007/978-3-030-75855-4_6.

[34] Hassija V, Chamola V, Mahapatra A, Singal A, Goel D, Huang K, et al. Interpreting black-box models: A review on explainable artificial intelligence. Cogn Comput 2023;16(1):45–74. http://dx.doi.org/10.1007/s12559-023-10179-8.

[35] Peters U. Explainable AI lacks regulative reasons: why AI and human decision-making are not equally opaque. AI Ethics 2022;3(3):963–74. http://dx.doi.org/10.1007/s43681-022-00217-w.

[36] Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D. A survey of methods for explaining black box models. ACM Comput Surv (CSUR) 2018;51(5):1–42.

[37] Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. 2017, arXiv:1702.08608. URL https://arxiv.org/abs/1702.08608.

[38] Holzinger A. Interactive machine learning for health informatics: when do we need the human-in-the-loop? Brain Inform 2016;3(2):119–31. http://dx.doi.org/10.1007/s40708-016-0042-6.

[39] Carvalho DV, Pereira EM, Cardoso JS. Machine learning interpretability: A survey on methods and metrics. Electronics 2019;8(8). http://dx.doi.org/10.3390/electronics8080832.

[40] Okolo CT, Aruleba K, Obaido G. Responsible AI in Africa—Challenges and opportunities. In: Social and cultural studies of robots and AI. Springer International Publishing; 2023, p. 35–64. http://dx.doi.org/10.1007/978-3-031-08215-3_3.

[41] Pearl J. Theoretical impediments to machine learning with seven sparks from the causal revolution. 2018, arXiv:1801.04016. URL https://arxiv.org/abs/1801.04016.

[42] Kumar S, Datta S, Singh V, Datta D, Kumar Singh S, Sharma R. Applications, challenges, and future directions of human-in-the-loop learning. IEEE Access 2024;12:75735–60. http://dx.doi.org/10.1109/ACCESS.2024.3401547.

[43] Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. ACM Comput Surv 2021;54(6). http://dx.doi.org/10.1145/3457607.

[44] Hong J-S, Lee J, Sim MK. Concise rule induction algorithm based on one-sided maximum decision tree approach. Expert Syst Appl 2024;237:121365. http://dx.doi.org/10.1016/j.eswa.2023.121365.

[45] Frnkranz J, Gamberger D, Lavrac N. Foundations of rule learning. Springer Publishing Company, Incorporated; 2012.

[46] Brown A, Tuor A, Hutchinson B, Nichols N. Recurrent neural network attention mechanisms for interpretable system log anomaly detection. In: Proceedings of the first workshop on machine learning for computing systems. 2018, p. 1–8.

[47] Dong Y, Su H, Zhu J, Zhang B. Improving interpretability of deep neural networks with semantic information. In: 2017 IEEE conference on computer vision and pattern recognition. 2017, p. 975–83. http://dx.doi.org/10.1109/CVPR.2017.110.

[48] Mienye ID, Swart TG. A hybrid deep learning approach with generative adversarial network for credit card fraud detection. Technologies 2024;12(10). http://dx.doi.org/10.3390/technologies12100186.

[49] Chaudhari S, Mithal V, Polatkan G, Ramanath R. An attentive survey of attention models. ACM Trans Intell Syst Technol 2021;12(5). http://dx.doi.org/10.1145/3465055.

[50] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. 2014, arXiv preprint arXiv:1409.0473.

[51] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: 2017 IEEE international conference on computer vision. 2017, p. 618–26. http://dx.doi.org/10.1109/ICCV.2017.74.

[52] Mienye E, Jere N, Obaido G, Mienye ID, Aruleba K. Deep learning in finance: A survey of applications and techniques. 2024, http://dx.doi.org/10.20944/preprints202408.1365.v1, Preprints.

[53] Kitson NK, Constantinou AC, Guo Z, Liu Y, Chobtham K. A survey of Bayesian network structure learning. Artif Intell Rev 2023;56(8):8721–814. http://dx.doi.org/10.1007/s10462-022-10351-w.

[54] Mienye ID, Jere N. Optimized ensemble learning approach with explainable AI for improved heart disease prediction. Information 2024;15(7). http://dx.doi.org/10.3390/info15070394, URL https://www.mdpi.com/2078-2489/15/7/394.

[55] Costa VG, Pedreira CE. Recent advances in decision trees: an updated survey. Artif Intell Rev 2022;56(5):4765–800. http://dx.doi.org/10.1007/s10462-022-10275-5.

[56] Mienye ID, Swart TG, Obaido G. Recurrent neural networks: A comprehensive review of architectures, variants, and applications. Information 2024;15(9):517. http://dx.doi.org/10.3390/info15090517.

[57] Ye Z, Yang W, Yang Y, Ouyang D. Interpretable machine learning methods for in vitro pharmaceutical formulation development. Food Front 2021;2(2):195–207. http://dx.doi.org/10.1002/fft2.78.

[58] Bifarin OO. Interpretable machine learning with tree-based shapley additive explanations: Application to metabolomics datasets for binary classification. In: Ashraf I, editor. PLOS ONE 2023;18(5):e0284315. http://dx.doi.org/10.1371/journal.pone.0284315.

[59] Dewi C, Tsai B-J, Chen R-C. Shapley additive explanations for text classification and sentiment analysis of internet movie database. In: Communications in computer and information science. Springer Nature Singapore; 2022, p. 69–80. http://dx.doi.org/10.1007/978-981-19-8234-7_6.

[60] Zhao X, Huang W, Huang X, Robu V, Flynn D. BayLIME: Bayesian local interpretable model-agnostic explanations. In: de Campos C, Maathuis MH, editors. Proceedings of the thirty-seventh conference on uncertainty in artificial intelligence. Proceedings of machine learning research, vol. 161, PMLR; 2021, p. 887–96, URL https://proceedings.mlr.press/v161/zhao21a.html.

[61] Ribeiro MT, Singh S, Guestrin C. "Why Should I Trust You?": Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. New York, NY, USA: Association for Computing Machinery; 2016, p. 1135–44. http://dx.doi.org/10.1145/2939672.2939778.

[62] Molnar C, Freiesleben T, König G, Herbinger J, Reisinger T, Casalicchio G, et al. Relating the partial dependence plot and permutation feature importance to the data generating process. In: Communications in computer and information science. Springer Nature Switzerland; 2023, p. 456–79. http://dx.doi.org/10.1007/978-3-031-44064-9_24.

[63] 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. J Med Syst 2021;45(5). http://dx.doi.org/10.1007/s10916-021-01736-5.

[64] Goldstein A, Kapelner A, Bleich J, Pitkin E. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. J Comput Graph Statist 2015;24(1):44–65. http://dx.doi.org/10.1080/10618600.2014.907095.

[65] Molnar C, König G, Bischl B, Casalicchio G. Model-agnostic feature importance and effects with dependent features: a conditional subgroup approach. Data Min Knowl Discov 2023;38(5):2903–41. http://dx.doi.org/10.1007/s10618-022-00901-9.

[66] Zhu X, Wang D, Pedrycz W, Li Z. Fuzzy rule-based local surrogate models for black-box model explanation. IEEE Trans Fuzzy Syst 2023;31(6):2056–64. http://dx.doi.org/10.1109/TFUZZ.2022.3218426.

[67] Ali M, Khattak AM, Ali Z, Hayat B, Idrees M, Pervez Z, et al. Estimation and interpretation of machine learning models with customized surrogate model. Electronics 2021;10(23). http://dx.doi.org/10.3390/electronics10233045, URL https://www.mdpi.com/2079-9292/10/23/3045.

[68] Del Ser J, Barredo-Arrieta A, Díaz-Rodríguez N, Herrera F, Saranti A, Holzinger A. On generating trustworthy counterfactual explanations. Inform Sci 2024;655:119898. http://dx.doi.org/10.1016/j.ins.2023.119898.

[69] Slack D, Hilgard A, Lakkaraju H, Singh S. Counterfactual explanations can be manipulated. In: Ranzato M, Beygelzimer A, Dauphin Y, Liang P, Vaughan JW, editors. In: Advances in neural information processing systems, vol. 34, Curran Associates, Inc.; 2021, p. 62–75.

[70] Hooker G, Mentch L, Zhou S. Unrestricted permutation forces extrapolation: variable importance requires at least one more model, or there is no free variable importance. Stat Comput 2021;31(6). http://dx.doi.org/10.1007/s11222-021-10057-z.

[71] Oh T, Kim D, Lee S, Won C, Kim S, Yang J-s, et al. Machine learning-based diagnosis and risk factor analysis of cardiocerebrovascular disease based on KNHANES. Sci Rep 2022;12(1). http://dx.doi.org/10.1038/s41598-022-06333-1.

[72] Velmurugan M, Ouyang C, Moreira C, Sindhgatta R. Evaluating fidelity of explainable methods for predictive process analytics. In: Lecture notes in business information processing. Springer International Publishing; 2021, p. 64–72. http://dx.doi.org/10.1007/978-3-030-79108-7_8.

[73] Miró-Nicolau M, i Capó AJ, Moyà-Alcover G. Assessing fidelity in XAI post-hoc techniques: A comparative study with ground truth explanations datasets. Artificial Intelligence 2024;335:104179. http://dx.doi.org/10.1016/j.artint.2024.104179.

[74] Visani G, Bagli E, Chesani F, Poluzzi A, Capuzzo D. Statistical stability indices for LIME: Obtaining reliable explanations for machine learning models. J Oper Res Soc 2021;73(1):91–101. http://dx.doi.org/10.1080/01605682.2020.1865846.

[75] Yeh C-K, Kim B, Arik S, Li C-L, Pfister T, Ravikumar P. On completeness-aware concept-based explanations in deep neural networks. In: Larochelle H, Ranzato M, Hadsell R, Balcan M, Lin H, editors. In: Advances in neural information processing systems, vol. 33, Curran Associates, Inc.; 2020, p. 20554–65.

[76] Zhou J, Gandomi AH, Chen F, Holzinger A. Evaluating the quality of machine learning explanations: A survey on methods and metrics. Electronics 2021;10(5). http://dx.doi.org/10.3390/electronics10050593, URL https://www.mdpi.com/2079-9292/10/5/593.

[77] Rudin C, Chen C, Chen Z, Huang H, Semenova L, Zhong C. Interpretable machine learning: Fundamental principles and 10 grand challenges. Stat Surv 2022;16:1–85.

[78] Goodman B, Flaxman S. European union regulations on algorithmic decision making and a "Right to Explanation". AI Mag 2017;38(3):50–7. http://dx.doi.org/10.1609/aimag.v38i3.2741.

[79] Woisetschläger H, Erben A, Marino B, Wang S, Lane ND, Mayer R, et al. Federated learning priorities under the European union artificial intelligence act. 2024, arXiv:2402.05968. URL https://arxiv.org/abs/2402.05968.

[80] Laux J, Wachter S, Mittelstadt B. Trustworthy artificial intelligence and the European union AI act: On the conflation of trustworthiness and acceptability of risk. Regul Gov 2023;18(1):3–32. http://dx.doi.org/10.1111/rego.12512.

[81] Mienye ID, Jere N. A survey of decision trees: Concepts, algorithms, and applications. IEEE Access 2024;12:86716–27. http://dx.doi.org/10.1109/ACCESS.2024.3416838.

[82] Mienye ID, Sun Y. Effective feature selection for improved prediction of heart disease. In: Lecture notes of the institute for computer sciences, social informatics and telecommunications engineering. Springer International Publishing; 2022, p. 94–107. http://dx.doi.org/10.1007/978-3-030-93314-2_6.

[83] Esmaeilzadeh P. Challenges and strategies for wide-scale artificial intelligence (AI) deployment in healthcare practices: A perspective for healthcare organizations. Artif Intell Med 2024;151:102861. http://dx.doi.org/10.1016/j.artmed.2024.102861, URL https://www.sciencedirect.com/science/article/pii/S0933365724001039.

[84] Xu T, White J, Kalkan S, Gunes H. Investigating bias and fairness in facial expression recognition. In: Lecture notes in computer science. Springer International Publishing; 2020, p. 506–23. http://dx.doi.org/10.1007/978-3-030-65414-6_35.

[85] Wang Y, Ma W, Zhang M, Liu Y, Ma S. A survey on the fairness of recommender systems. ACM Trans Inf Syst 2023;41(3):1–43. http://dx.doi.org/10.1145/3547333.

[86] Islam R, Keya KN, Pan S, Sarwate AD, Foulds JR. Differential fairness: An intersectional framework for fair AI. Entropy 2023;25(4). http://dx.doi.org/10.3390/e25040660, URL https://www.mdpi.com/1099-4300/25/4/660.

[87] De Schutter L, De Cremer D. How counterfactual fairness modelling in algorithms can promote ethical decision-making. Int J Hum–Comput Interact 2023;40(1):33–44. http://dx.doi.org/10.1080/10447318.2023.2247624.

[88] Harris CG. Combining human-in-the-loop systems and AI fairness toolkits to reduce age bias in AI job hiring algorithms. In: 2024 IEEE international conference on big data and smart computing. 2024, p. 60–6. http://dx.doi.org/10.1109/BigComp60711.2024.00019.

[89] Beneyto A, Puig V, Bequette BW, Vehi J. A hybrid automata approach for monitoring the patient in the loop in artificial pancreas systems. Sensors 2021;21(21). http://dx.doi.org/10.3390/s21217117, URL https://www.mdpi.com/1424-8220/21/21/7117.

[90] Retzlaff CO, Das S, Wayllace C, Mousavi P, Afshari M, Yang T, et al. Human-in-the-loop reinforcement learning: A survey and position on requirements, challenges, and opportunities. J Artificial Intelligence Res 2024;79:359–415. http://dx.doi.org/10.1613/jair.1.15348.

[91] Cashaback JGA, Allen JL, Chou AH-Y, Lin DJ, Price MA, Secerovic NK, et al. NSF DARE—transforming modeling in neurorehabilitation: a patient-in-the-loop framework. J NeuroEng Rehabil 2024;21(1). http://dx.doi.org/10.1186/s12984-024-01318-9.

[92] Holzinger A, Langs G, Denk H, Zatloukal K, Müller H. Causability and explainability of artificial intelligence in medicine. WIREs Data Min Knowl Discov 2019;9(4). http://dx.doi.org/10.1002/widm.1312.

[93] Tjoa E, Guan C. A survey on explainable artificial intelligence (XAI): Toward medical XAI. IEEE Trans Neural Netw Learn Syst 2021;32(11):4793–813. http://dx.doi.org/10.1109/TNNLS.2020.3027314.

[94] Hase P, Chen C, Li O, Rudin C. Interpretable image recognition with hierarchical prototypes. In: Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, In: Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, 2019;vol. 7:32–40.http://dx.doi.org/10.1609/hcomp.v7i1.5265,

[95] Lundberg S, Lee S-I. A unified approach to interpreting model predictions. 2017, arXiv:1705.07874. URL https://arxiv.org/abs/1705.07874.

[96] Li X, Wu R, Zhao W, Shi R, Zhu Y, Wang Z, et al. Machine learning algorithm to predict mortality in critically ill patients with sepsis-associated acute kidney injury. Sci Rep 2023;13(1). http://dx.doi.org/10.1038/s41598-023-32160-z.

[97] Tonekaboni S, Joshi S, McCradden MD, Goldenberg A. What clinicians want: Contextualizing explainable machine learning for clinical end use. In: Doshi-Velez F, Fackler J, Jung K, Kale D, Ranganath R, Wallace B, et al., editors. Proceedings of the 4th machine learning for healthcare conference. Proceedings of machine learning research, vol. 106, PMLR; 2019, p. 359–80, URL https://proceedings.mlr.press/v106/tonekaboni19a.html.

[98] Bento V, Kohler M, Diaz P, Mendoza L, Pacheco MA. Improving deep learning performance by using explainable artificial intelligence (XAI) approaches. Discov Artif Intell 2021;1(1). http://dx.doi.org/10.1007/s44163-021-00008-y.

[99] Ou S-M, Tsai M-T, Lee K-H, Tseng W-C, Yang C-Y, Chen T-H, et al. 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(1). http://dx.doi.org/10.1186/s13040-023-00324-2.

[100] Alkhamis MA, Al Jarallah M, Attur S, Zubaid M. Interpretable machine learning models for predicting in-hospital and 30 days adverse events in acute coronary syndrome patients in Kuwait. Sci Rep 2024;14(1):1243.

[101] Thiruvenkadam K, Ravindran V, Thiyagarajan A. Deep learning with XAI based multi-modal MRI brain tumor image analysis using image fusion techniques. In: 2024 international conference on trends in quantum computing and emerging business technologies. 2024, p. 1–5. http://dx.doi.org/10.1109/TQCEBT59414.2024.10545215.

[102] Sirapangi MD, Gopikrishnan S. MAIPFE: An efficient multimodal approach integrating pre-emptive analysis, personalized feature selection, and explainable AI. Comput Mater Contin 2024;79(2):2229–51. http://dx.doi.org/10.32604/cmc.2024.047438.

[103] Khater T, Tawfik H, Singh B. Explainable artificial intelligence for investigating the effect of lifestyle factors on obesity. Intell Syst Appl 2024;23:200427. http://dx.doi.org/10.1016/j.iswa.2024.200427, URL https://www.sciencedirect.com/science/article/pii/S2667305324001017.

[104] Mienye ID, Kenneth Ainah P, Emmanuel ID, Esenogho E. Sparse noise minimization in image classification using genetic algorithm and DenseNet. In: 2021 conference on information communications technology and society. 2021, p. 103–8. http://dx.doi.org/10.1109/ICTAS50802.2021.9395014.

[105] Salehi AW, Khan S, Gupta G, Alabduallah BI, Almjally A, Alsolai H, et al. A study of CNN and transfer learning in medical imaging: Advantages, challenges, future scope. Sustainability 2023;15(7). http://dx.doi.org/10.3390/su15075930, URL https://www.mdpi.com/2071-1050/15/7/5930.

[106] Ahmed M, Bibi T, Khan RA, Nasir S. Enhancing breast cancer diagnosis in mammography: Evaluation and integration of convolutional neural networks and explainable AI. 2024, arXiv:2404.03892. URL https://arxiv.org/abs/2404.03892.

[107] DeGrave AJ, Janizek JD, Lee S-I. AI for radiographic COVID-19 detection selects shortcuts over signal. Nat Mach Intell 2021;3(7):610–9. http://dx.doi.org/10.1038/s42256-021-00338-7.

[108] Baumgartner CF, Koch LM, Tezcan KC, Ang JX, Konukoglu E. Visual feature attribution using wasserstein GANs. 2018, arXiv:1711.08998. URL https://arxiv.org/abs/1711.08998.

[109] Gargeya R, Leng T. Automated identification of diabetic retinopathy using deep learning. Ophthalmology 2017;124(7):962–9. http://dx.doi.org/10.1016/j.ophtha.2017.02.008, URL https://www.sciencedirect.com/science/article/pii/S0161642016317742.

[110] Musthafa MM, Mahesh TR, Kumar VV, Guluwadi S. Enhancing brain tumor detection in MRI images through explainable AI using grad-CAM with resnet 50. BMC Med Imaging 2024;24(1). http://dx.doi.org/10.1186/s12880-024-01292-7.

[111] Wang X, Wang W, Ren H, Li X, Wen Y. Prediction and analysis of risk factors for diabetic retinopathy based on machine learning and interpretable models. Heliyon 2024;10(9):e29497. http://dx.doi.org/10.1016/j.heliyon.2024.e29497.

[112] Albahri A, Duhaim AM, Fadhel MA, Alnoor A, Baqer NS, Alzubaidi L, et al. A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion. Inf Fusion 2023;96:156–91. http://dx.doi.org/10.1016/j.inffus.2023.03.008, URL https://www.sciencedirect.com/science/article/pii/S1566253523000891.

[113] Ramoliya F, Gohil K, Gohil A, Gupta R, Kakkar R, Tanwar S, et al. X-CaD: Explainable AI for skin cancer diagnosis in healthcare 4.0 telesurgery. In: ICC 2024 - IEEE international conference on communications. 2024, p. 238–43. http://dx.doi.org/10.1109/ICC51166.2024.10622832.

[114] Arcadu F, Benmansour F, Maunz A, Willis J, Haskova Z, Prunotto M. Deep learning algorithm predicts diabetic retinopathy progression in individual patients. npj Digit Med 2019;2(1). http://dx.doi.org/10.1038/s41746-019-0172-3.

[115] Patel M, Gohil K, Gohil A, Ramoliya F, Gupta R, Tanwar S, et al. Explainable AI for gastrointestinal disease diagnosis in telesurgery healthcare 4.0. Comput Electr Eng 2024;118:109414. http://dx.doi.org/10.1016/j.compeleceng.2024.109414, URL https://www.sciencedirect.com/science/article/pii/S0045790624003422.

[116] Awotunde JB, Jimoh RG, Adeniyi AE, Ayo EF, Ajamu GJ, Aremu DR. Application of interpretable artificial intelligence enabled cognitive internet of things for COVID-19 pandemics. In: Explainable machine learning for multimedia based healthcare applications. Springer International Publishing; 2023, p. 191–213. http://dx.doi.org/10.1007/978-3-031-38036-5_11.

[117] De Fauw J, Ledsam JR, Romera-Paredes B, Nikolov S, Tomasev N, Blackwell S, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. Nature Med 2018;24(9):1342–50. http://dx.doi.org/10.1038/s41591-018-0107-6.

[118] Ribeiro MM, Rocha LB, Leal CC, Santos AR, Pires DS, Santana E, et al. Automatic detection of arrhythmias from imbalanced data using machine learning techniques. Expert Syst Appl 2020;158:113551.

[119] Li H, Yang Y, Hong W, Huang M, Wu M, Zhao X. Applications of genome editing technology in the targeted therapy of human diseases: mechanisms, advances and prospects. Signal Transduct Target Ther 2020;5(1). http://dx.doi.org/10.1038/s41392-019-0089-y.

[120] Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. Brief Bioinform 2018;19(6):1236–46.

[121] Hauser K, Kurz A, Haggenmüller S, Maron RC, von Kalle C, Utikal JS, et al. Explainable artificial intelligence in skin cancer recognition: A systematic review. Eur J Cancer 2022;167:54–69. http://dx.doi.org/10.1016/j.ejca.2022.02.025, URL https://www.sciencedirect.com/science/article/pii/S095980492200123X.

[122] Char DS, Shah NH, Magnus D. Implementing machine learning in health care — Addressing ethical challenges. N Engl J Med 2018;378(11):981–3. http://dx.doi.org/10.1056/nejmp1714229.

[123] Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. Nat Mach Intell 2019;1(9):389–99. http://dx.doi.org/10.1038/s42256-019-0088-2.

[124] Garrido-Muñoz I, Montejo-Ráez A, Martínez-Santiago F, Ureña-López LA. A survey on bias in deep NLP. Appl Sci 2021;11(7). http://dx.doi.org/10.3390/app11073184, URL https://www.mdpi.com/2076-3417/11/7/3184.

[125] Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring fairness in machine learning to advance health equity. Ann Intern Med 2018;169(12):866–72. http://dx.doi.org/10.7326/M18-1990, PMID: 30508424. URL https://www.acpjournals.org/doi/abs/10.7326/M18-1990.

[126] Mitchell S, Potash E, Barocas S, D'Amour A, Lum K. Algorithmic fairness: Choices, assumptions, and definitions. Annu Rev Stat Appl 2021;8(1):141–63. http://dx.doi.org/10.1146/annurev-statistics-042720-125902.

[127] He J, Hao Y, Wang X. An interpretable aid decision-making model for flag state control ship detention based on SMOTE and XGBoost. J Mar Sci Eng 2021;9(2). http://dx.doi.org/10.3390/jmse9020156, URL https://www.mdpi.com/2077-1312/9/2/156.

[128] Lipton ZC. The mythos of model interpretability. Commun ACM 2018;61(10):36–43. http://dx.doi.org/10.1145/3233231.

[129] Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 2019;1(5):206–15. http://dx.doi.org/10.1038/s42256-019-0048-x.

[130] Yang Z, Zhang A, Sudjianto A. GAMI-Net: An explainable neural network based on generalized additive models with structured interactions. Pattern Recognit 2021;120:108192. http://dx.doi.org/10.1016/j.patcog.2021.108192, URL https://www.sciencedirect.com/science/article/pii/S0031320321003484.

[131] Kuang K, Li L, Geng Z, Xu L, Zhang K, Liao B, et al. Causal inference. Engineering 2020;6(3):253–63. http://dx.doi.org/10.1016/j.eng.2019.08.016, URL https://www.sciencedirect.com/science/article/pii/S2095809919305235.