Review

Bridging the Gap Between Black Box AI and Clinical Practice: Advancing Explainable AI for Trust, Ethics, and Personalized Healthcare Diagnostics

Dang Anh Tuan *

Posted Date: 27 September 2024

doi: 10.20944/preprints202409.1974.v2

Keywords: Explainable AI (XAI); Healthcare diagnostics; Grad-CAM; SHAP; LIME; Trust in AI; Ethical AI; Bias mitigation; Regulatory compliance; Personalized care; AI interpretability; Clinical decision-making

Review

Bridging the Gap Between Black Box AI and Clinical Practice: Advancing Explainable AI for Trust, Ethics, and Personalized Healthcare Diagnostics

Dang Anh Tuan

Faculty of Information Technology, Industrial University of Ho Chi Minh City, Ho Chi Minh City, 70000, Vietnam; tuanda222@gmail.com; Tel: +84987 066 222

Abstract: Explainable AI (XAI) has emerged as a pivotal tool in healthcare diagnostics, offering much-needed transparency and interpretability in complex AI models. XAI techniques, such as SHAP, Grad-CAM, and LIME, enable clinicians to understand AI-driven decisions, fostering greater trust and collaboration between human and machine in clinical settings. This review explores the key benefits of XAI in enhancing diagnostic accuracy, personalizing patient care, and ensuring compliance with regulatory standards. However, despite its advantages, XAI faces significant challenges, including balancing model accuracy with interpretability, scaling for real-time clinical use, and mitigating biases inherent in medical data. Ethical concerns, particularly surrounding fairness and accountability, are also discussed in relation to AI's growing role in healthcare. The review emphasizes the importance of developing hybrid models that combine high accuracy with improved interpretability and suggests that future research should focus on explainable-by-design systems, reducing computational costs, and addressing ethical issues. As AI continues to integrate into healthcare, XAI will play an essential role in ensuring that AI systems are transparent, accountable, and aligned with the ethical standards required in clinical practice.

Keywords: Explainable AI (XAI); Healthcare diagnostics; Grad-CAM; SHAP; LIME; Trust in AI; Ethical AI; Bias mitigation; Regulatory compliance; Personalized care; AI interpretability; Clinical decision-making

1. Introduction

1.1. Background and Motivation

Artificial Intelligence (AI) has become an indispensable tool in modern healthcare, revolutionizing various areas, particularly diagnostics. AI systems, especially machine learning (ML) and deep learning (DL) models, have shown remarkable success in automating the detection of diseases, predicting patient outcomes, and analyzing complex medical data, such as medical images and genomic sequences. For instance, convolutional neural networks (CNNs) have been widely used for detecting abnormalities in medical images, including tumor detection in MRI and CT scans, outperforming traditional diagnostic methods in both speed and accuracy [1]. Moreover, AI's ability to analyze big data has led to breakthroughs in personalized medicine and predictive analytics for patient outcomes, opening new avenues for precision healthcare [2].

However, the adoption of AI in clinical practice faces a critical challenge: the "black-box" nature of many AI models. While these models deliver highly accurate predictions, they often do so without providing clear explanations of how or why a particular decision was made. This opacity poses a significant issue in healthcare, where transparency is essential for clinical decision-making. Physicians need to understand AI's rationale to trust its recommendations and to communicate them effectively to patients. Moreover, healthcare decisions frequently carry life-or-death implications, and unexplained AI decisions could lead to a lack of accountability and increased legal risks [3].

This is where Explainable AI (XAI) comes into play. XAI seeks to provide clarity by making AI models more interpretable and transparent. The concept of XAI is crucial for bridging the gap

doi:10.20944/preprints202409.1974.v1

between AI's decision-making processes and clinical needs, enabling healthcare professionals to understand and trust AI systems more fully. This is not just a technical challenge but also an ethical one, as explainability is fundamental for ensuring fairness and accountability in medical AI applications [4].

1.2. Scope and Objectives of the Review

The objective of this review is to provide a comprehensive analysis of the current state of XAI in healthcare diagnostics. We will examine the leading XAI techniques, including post-hoc explainability methods such as LIME and SHAP, as well as inherently interpretable models like decision trees and rule-based systems. The review will focus on how these methods are being applied to critical areas such as medical imaging, genomics, and predictive analytics for patient outcomes. Furthermore, we will explore the trade-offs between explainability and performance in AI models, discuss the ethical and regulatory implications, and identify the barriers to the widespread adoption of XAI in clinical practice.

In this review, we aim to address the key question: How can XAI improve the transparency and trustworthiness of AI systems in healthcare, and what are the remaining challenges to its full integration into medical practice? By investigating these issues, we hope to offer valuable insights for future research and development in this rapidly evolving field, ensuring that AI becomes not only an accurate but also an interpretable and reliable tool in healthcare diagnostics.

2. Overview of Artificial Intelligence in Healthcare Diagnostics

2.1. AI Techniques in Healthcare

AI has revolutionized healthcare diagnostics by introducing computational models that mimic human cognitive functions, primarily through ML and DL techniques. These methods have become indispensable in analyzing large, complex datasets, enabling precise diagnosis, prognosis, and treatment personalization. Among the most prominent AI techniques used in healthcare diagnostics are convolutional neural networks (CNNs) [5–7], recurrent neural networks (RNNs) [8], decision trees [7,9], and support vector machines (SVMs) [7,8,10], each with its unique advantages and applications.

CNNs have become the gold standard for analyzing medical images due to their ability to automatically detect patterns and extract relevant features from data, often surpassing traditional image analysis methods. CNNs have been widely used in the detection of tumors, brain anomalies, and other pathologies from modalities such as magnetic resonance imaging (MRI), computed tomography (CT), and X-rays [11–15]. For instance, CNNs have been instrumental in automating breast cancer detection, significantly improving diagnostic accuracy and reducing human error in mammogram analysis [16–21].

RNNs and their variant, long short-term memory (LSTM) networks, have been particularly effective in processing sequential medical data, such as time-series data from electrocardiograms (ECGs) or patient health records. RNNs excel in capturing temporal dependencies, making them valuable for predicting disease progression or patient outcomes based on historical data [22–27]. LSTM models have been applied in predictive healthcare analytics, such as predicting heart failure readmissions or monitoring patients with chronic diseases [25,26,28–33].

In addition to deep learning techniques, traditional machine learning methods like decision trees and SVMs continue to play a role in healthcare diagnostics, especially in scenarios where interpretable models are required. Decision trees are favored for their simplicity and transparency, allowing healthcare professionals to understand the logic behind AI-driven decisions. SVMs, on the other hand, are powerful classifiers used for tasks such as gene expression analysis and early disease detection in oncology [6,34–38].

The most commonly used AI techniques in healthcare diagnostics are CNNs, RNNs, SVMs, and Decision Trees, each with specific applications and trade-offs. Table 1 provides a comparative summary of these techniques in terms of their healthcare applications, advantages, and limitations.

Table 1. Comparison of AI Techniques in Healthcare Diagnostics.

Technique	Application in Healthcare	Advantages	Limitations	References
CNN	Medical imaging (e.g., MRI, CT, X-rays)	High accuracy, automatic feature extraction	Black-box nature, requires large datasets	[6,38–40]
RNN	Time-series data (e.g., ECGs, patient history)	Captures temporal dependencies in medical data	Difficult to train, prone to vanishing gradient	[5,41,42]
SVM	Cancer detection, gene expression analysis	Effective in small, high-dimensional datasets	Less effective for large, unstructured data	[39,43,44]
Decision Trees	Diagnosis support in clinical data	Easy to interpret, transparent decision-making	Prone to overfitting, lower accuracy compared to DL	[5,38,45]

Table 1 shows CNNs are primarily used in medical imaging due to their high accuracy in automatic feature extraction from large image datasets. However, their black-box nature makes them difficult to interpret [46,47]. RNNs are suitable for time-series data, such as patient histories or ECGs, but they are difficult to train and can suffer from vanishing gradient problems [48]. SVMs are highly effective in small, high-dimensional datasets, making them a strong choice for tasks like gene expression analysis, although they struggle with unstructured data [49,50]. Decision trees provide transparency and ease of interpretation but tend to overfit, especially in more complex healthcare datasets [51].

2.2. Challenges of Black-box AI in Healthcare

Despite the successes of AI in healthcare, the widespread adoption of AI systems in clinical diagnostics is hindered by several challenges, with the most prominent being the "black-box" nature of many advanced AI models, particularly deep learning. Black-box models, such as deep neural networks, are typically highly complex, comprising multiple layers of computation that make their decision-making processes opaque to users, including medical professionals. This lack of transparency poses significant risks in the healthcare domain, where clinical decisions must be interpretable, justifiable, and accountable [5,52–54].

In clinical practice, trust is a critical component of decision-making. Physicians and healthcare providers need to understand and justify the decisions suggested by AI models, especially in critical diagnoses where lives are at stake. The inability to explain how or why a model arrived at a particular diagnosis or prognosis can result in hesitancy among clinicians to adopt AI tools, even if these tools are more accurate than human judgment in some cases [55–57]. This lack of transparency not only undermines trust but can also have legal and ethical implications, particularly if a model's decision leads to an adverse outcome [58].

Moreover, healthcare is a highly regulated field, with stringent legal and ethical standards requiring full accountability for clinical decisions. When AI models provide no clear rationale for their outputs, it becomes challenging to validate or scrutinize their decisions in a legal or regulatory framework. For example, if an AI system recommends a particular treatment but fails to provide a comprehensible explanation for this recommendation, it could be difficult for clinicians to defend that decision in court if something goes wrong [55,59–61]. Furthermore, AI systems trained on biased or incomplete datasets can introduce unintended biases into medical decision-making, potentially leading to discriminatory outcomes, particularly for minority or underrepresented groups [62–68].

The inherent opacity of many AI models also complicates their integration into existing clinical workflows. Healthcare professionals often prefer models that provide interpretable and actionable insights, allowing them to incorporate AI recommendations into their decision-making processes confidently. However, black-box models typically offer little to no insight into their decision pathways, which contrasts with the need for clear, evidence-based reasoning in medicine [69].

In summary, while AI holds tremendous potential for revolutionizing healthcare diagnostics, its black-box nature remains a significant barrier to adoption. The need for interpretability, transparency, and accountability in AI-driven healthcare is paramount to ensuring that AI systems are trusted and integrated into clinical practice effectively. This challenge has given rise to the development of XAI, a set of methods aimed at making AI models more understandable and interpretable, which will be explored further in the next sections of this review.

Grad-CAM is a widely used method to provide visual explanations for CNNs in medical imaging. By applying Grad-CAM, clinicians can see the regions of medical images that the model has focused on when making a diagnostic decision [70]. As shown in **Figure 1**, the heatmap highlights the areas of importance in a brain MRI for detecting a tumor.

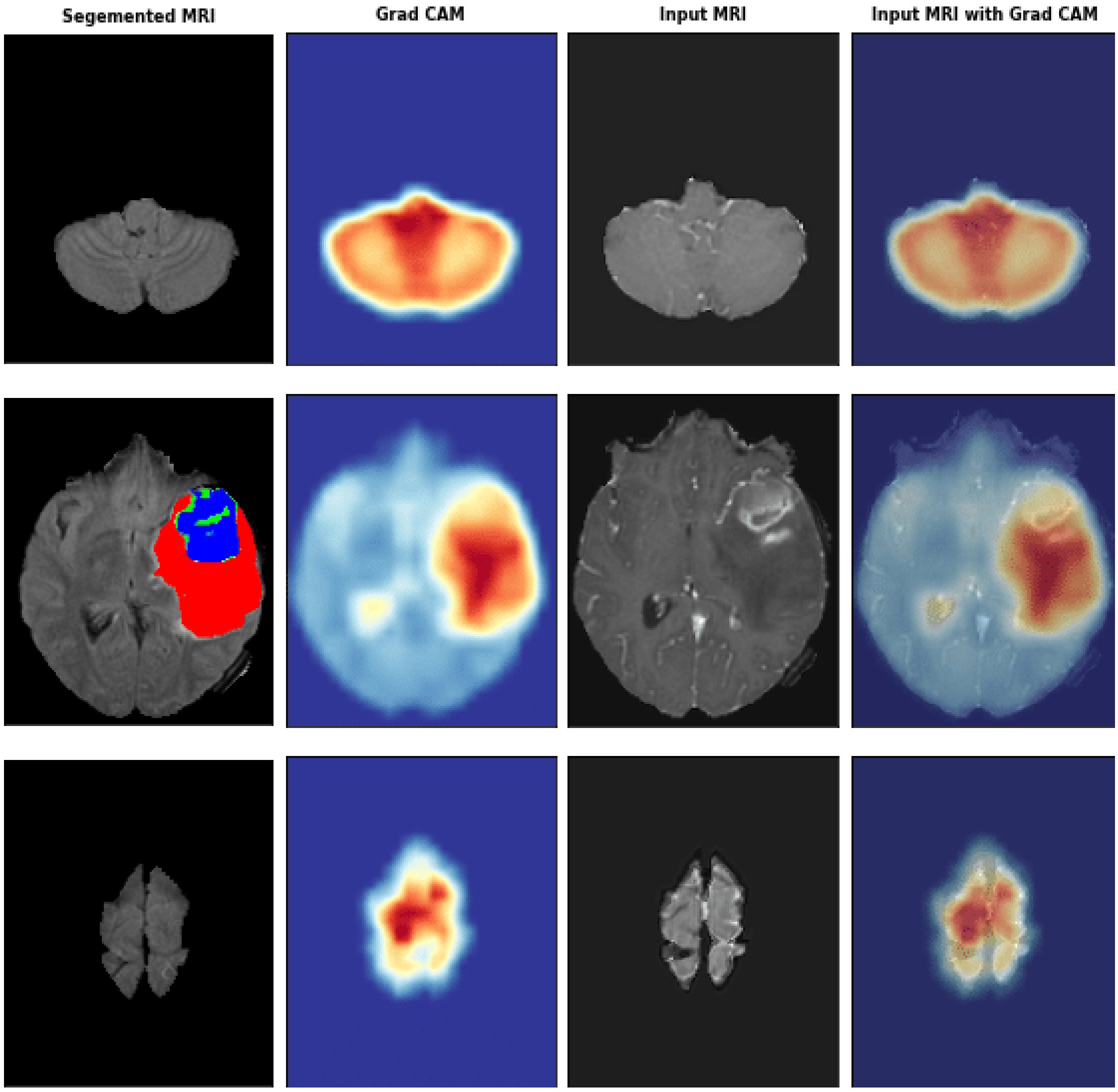


Figure 1. Grad-CAM visualizations generated at different depths of the same MRI volume. Row 2 highlights regions where the tumor exists, while Row 1 shows regions without the tumor from a bottom view, and Row 3 depicts regions without the tumor from a top view of the brain. The heatmap

indicates areas of importance in the CNN's analysis, with warmer colors (red, orange) showing regions of high significance for tumor detection.

3. Explainable AI: Concepts and Techniques

3.1. Definition of Explainability

XAI is an evolving subfield within artificial intelligence that addresses the need to make machine learning models, particularly complex black-box models like deep neural networks, more transparent and interpretable. In the context of healthcare, where diagnostic decisions may have life-or-death consequences, explainability is critical for both ethical and practical reasons. Medical professionals require a clear understanding of how an AI model arrives at a specific diagnosis or prediction to ensure its reliability and to gain clinical trust in the system [4].

Explainability can be categorized into three main levels:

Model-Level Explainability: Provides a global view of how a model functions as a whole, such as decision trees or linear regression models, where decision paths and feature importances are inherently interpretable [71–74].

Decision-Level Explainability: Focuses on explaining specific predictions, such as why a model classified a particular scan as indicative of a tumor. Techniques like SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) are used to achieve decision-level explanations [75].

Process-Level Explainability: Involves understanding how the AI model processes input data through different layers or components, a common practice in deep learning where visualization techniques like Grad-CAM (Gradient-weighted Class Activation Mapping) highlight relevant image regions that contribute to predictions [76].

With growing concerns about AI's "black-box" nature, XAI techniques have gained significant attention in healthcare diagnostics, where the ability to interpret and verify AI outputs is crucial for clinical validation and compliance with regulatory requirements.

These varying levels of explainability address different needs within healthcare diagnostics. For example, model-level explainability helps healthcare institutions validate AI systems before deployment by offering a global understanding of how the model functions, while decision-level explainability is particularly valuable for clinicians who need to understand the rationale behind individual predictions. Process-level explainability, on the other hand, provides insights into how deep learning models handle and process complex medical data, often giving clinicians visual cues to validate AI-driven diagnoses.

Given the critical role of explainability in healthcare, various XAI techniques have been developed, each with its own strengths and limitations. Some methods, like SHAP and LIME, excel at providing local explanations for individual predictions, while others, like Grad-CAM, are more suited to interpreting deep learning models, especially in image-based diagnostics. Table 2 provides a comparative overview of the most widely used XAI techniques in healthcare, summarizing their key features, advantages, and limitations. This comparison helps clinicians and researchers select the most appropriate method depending on the model architecture and clinical requirements.

Table 2. Comparison of Common Explainable AI Techniques in Healthcare.

Technique	Type	Explanation Process	Application in Healthcare	Limitations	References
LIME	Post-hoc	Locally perturbs input data to approximate the decision	Useful in explaining complex models for medical	Only provides local interpretability; less effective for	[77–79]

		boundary of the model around the instance being explained.	image classification or patient diagnosis.	very large datasets.	
SHAP	Post-hoc	Based on cooperative game theory, assigns Shapley values to features that contributed to a prediction.	Often applied in risk prediction models for chronic diseases, patient readmission, or treatment planning.	Computationally expensive for large models.	[80–82]
Grab-CAM	Post-hoc	Highlights the regions of an image that are most relevant to the model's predictions using heatmaps.	Widely used in medical imaging for identifying regions of interest, such as tumors.	Limited to convolutional neural networks.	[79,83,84]
DeepLIFT	Post-hoc	Tracks the contribution of each input feature relative to a reference input, improving gradient-based methods.	Applied in genomics and precision medicine for feature attribution.	Less interpretable for complex temporal models.	[53,77,81]
Decision Trees	Inherently Interpretable	Visualizes decision-making through a tree of logical	Effective in rule-based diagnosis and decision	Prone to overfitting and less accurate in comparison to deep models.	[81,82,85]

		conditions, making predictions transparent.	support systems.		

3.2. Common XAI Techniques

XAI techniques fall broadly into two categories: post-hoc explainability methods, which are applied after a model has been trained, and inherently interpretable models, which are designed to be transparent from the start. In healthcare diagnostics, post-hoc methods are commonly used to explain predictions from high-performing but opaque models like CNNs.

Post-hoc Explainability Methods:

LIME: LIME works by perturbing the input data (e.g., changing pixels in an image) and observing how these changes impact the model's prediction. This method is widely used in interpreting image classification models and personalized medicine to explain individual diagnoses [86–88]. However, it is limited by its local nature, providing insight into a model’s behavior only around a specific prediction rather than globally [89].

SHAP: SHAP is grounded in cooperative game theory and calculates the Shapley value for each feature, offering a global perspective on feature importance across all predictions. SHAP is highly valuable in healthcare applications like risk prediction for chronic diseases or feature attribution in genomic studies [90]. The downside is its computational complexity, especially for deep learning models [91–94].

Grad-CAM: Grad-CAM is particularly useful in medical imaging, where it overlays heatmaps on input images to visualize which regions most influence the CNN’s decision [95]. For example, it can highlight tumor regions in MRI and CT scans, making the CNN’s reasoning transparent [96–98].

DeepLIFT: This method is an improvement on gradient-based techniques, calculating the contribution of each input feature relative to a baseline reference. It has been particularly useful in genomics, where it helps explain the influence of specific genes or sequences on a model’s output [99].

Inherently Interpretable Models:

Decision Trees: Decision trees remain popular in healthcare diagnostics for their simplicity and transparency, where each decision is based on clear, logical steps. They are used in decision support systems, such as diagnosing diseases based on clinical guidelines. However, decision trees often suffer from overfitting and lower accuracy compared to more complex models like deep neural networks [100–104].

doi:10.20944/preprints202409.1974.v1

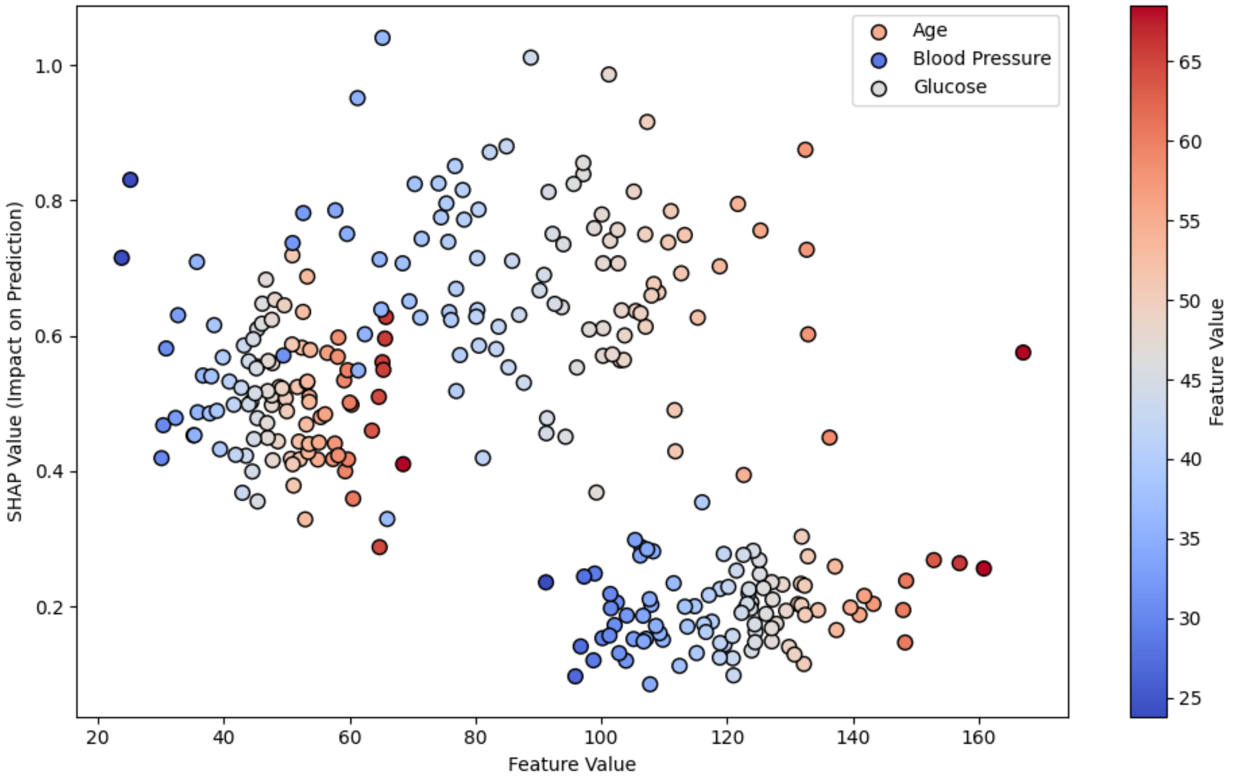


Figure 2. SHAP Values for Risk Prediction in Chronic Disease Diagnosis. The figure illustrates SHAP values for three key features—age, blood pressure, and glucose levels—in a chronic disease risk prediction model. Each dot represents a patient, with the color indicating the feature's value (e.g., red for higher values and blue for lower values). The x-axis shows the individual feature values, while the y-axis represents the corresponding SHAP values, which measure the impact of each feature on the model's prediction. Positive SHAP values indicate an increased likelihood of disease diagnosis, while negative values suggest a decreased likelihood. This visualization aids in interpreting how the model weighs different features, providing transparency in clinical decision-making.

The Table 2 compares different XAI techniques, illustrating their respective applications in healthcare diagnostics. In particular, post-hoc methods like SHAP and LIME provide detailed explanations of individual predictions, while Grad-CAM is more suited to image-based tasks. Figure 2 demonstrates SHAP values in a risk prediction model for chronic diseases, illustrating how each feature contributes to the final prediction. This type of visual explanation is crucial for understanding how an AI model evaluates various risk factors, making the model's decision-making process more transparent to healthcare professionals.

4. Applications of XAI in Healthcare Diagnostics

XAI continues to make significant contributions to healthcare diagnostics by increasing the transparency, trust, and utility of AI systems. From improving the interpretability of deep learning models in medical imaging to supporting personalized medicine and chronic disease risk prediction, XAI is crucial in bridging the gap between black-box AI models and clinical decision-making. Table 3 provides an overview of how different XAI techniques are applied in various healthcare domains, outlining their specific purposes and example use cases that demonstrate their impact.

4.1. Medical Imaging

One of the key areas where XAI has made a profound impact is in medical imaging. Complex AI models, such as CNNs, have revolutionized diagnostic imaging by detecting anomalies like tumors, fractures, and organ irregularities. However, the "black-box" nature of these models presents challenges in clinical adoption, as clinicians need transparency to trust the system's predictions [105].

Grad-CAM (Gradient-weighted Class Activation Mapping) remains a widely used XAI method in medical imaging, allowing practitioners to see the specific regions of an image that influence the AI's predictions. Recent advancements in ensemble models combining transfer learning with vision transformers have further enhanced the diagnostic accuracy and explainability of models, especially in Alzheimer's disease diagnosis [106,107]. In this approach, a combination of transfer learning (using models like ResNet50 and DenseNet121) with explainability techniques has led to higher accuracy and interpretability, improving diagnostic confidence [107].

Example Use Case: In Alzheimer's diagnosis, Grad-CAM is used to highlight brain regions in MRI scans that are most significant for the model's prediction. This transparency allows neurologists to verify whether the AI's focus corresponds to clinically relevant areas, improving trust in AI-driven diagnostic tools [106,107].

4.2. Personalized Medicine in Oncology

XAI is also becoming essential in personalized medicine, particularly in oncology. Personalized treatment plans based on a patient's genetic makeup, tumor characteristics, and clinical history are increasingly being developed using AI models. However, the success of such models depends on their ability to provide clear and understandable treatment recommendations to healthcare providers.

A recent 2024 study used XAI-empowered decision trees for predicting personalized breast cancer treatment options, including hormonal therapies, chemotherapy, and anti-HER2 treatments. The model achieved 99.87% accuracy and improved transparency by explaining which clinical and genetic factors influenced the treatment recommendation [108]. This approach is key to building trust between clinicians and AI systems, as it allows for an understandable rationale behind treatment decisions [109].

Example Use Case: For breast cancer patients, an AI model using clinical and genomic data can recommend specific therapies (e.g., anti-HER2 therapy) based on a decision tree model. With XAI, clinicians can see which genetic mutations or clinical factors led to this recommendation, ensuring that treatment aligns with the patient's specific profile [108].

4.3. Chronic Disease Risk Prediction

XAI techniques are widely used in predictive analytics for chronic diseases such as diabetes, heart failure, and hypertension. SHAP is particularly useful for identifying the most critical risk factors in predictive models, providing clear explanations for individual patient outcomes. For instance, in predicting heart failure readmissions, SHAP values help explain which factors—such as previous hospitalizations, blood pressure levels, or medication adherence—are driving the model's predictions [110].

A comprehensive review of AI-based healthcare techniques from 2024 highlights the application of SHAP in identifying key risk factors in disease comorbidity and predicting future health events. XAI techniques like SHAP have allowed clinicians to better understand and intervene in high-risk cases, offering more tailored preventive care [39].

Example Use Case: In predicting heart failure readmission, SHAP values indicate that past hospitalizations and blood pressure levels are the most significant contributors to the model's prediction. By understanding these factors, clinicians can prioritize preventive measures, such as medication adjustments, to reduce the risk of readmission.

4.4. Natural Language Processing (NLP) in Healthcare

Recent research has also focused on Explainable and Interpretable AI (XIAI) in natural language processing (NLP) for healthcare applications. As large language models (LLMs) such as GPT-3 and

doi:10.20944/preprints202409.1974.v

BERT are increasingly applied to medical tasks—such as analyzing electronic health records (EHRs) or generating diagnostic reports—the need for explainability has become more critical.

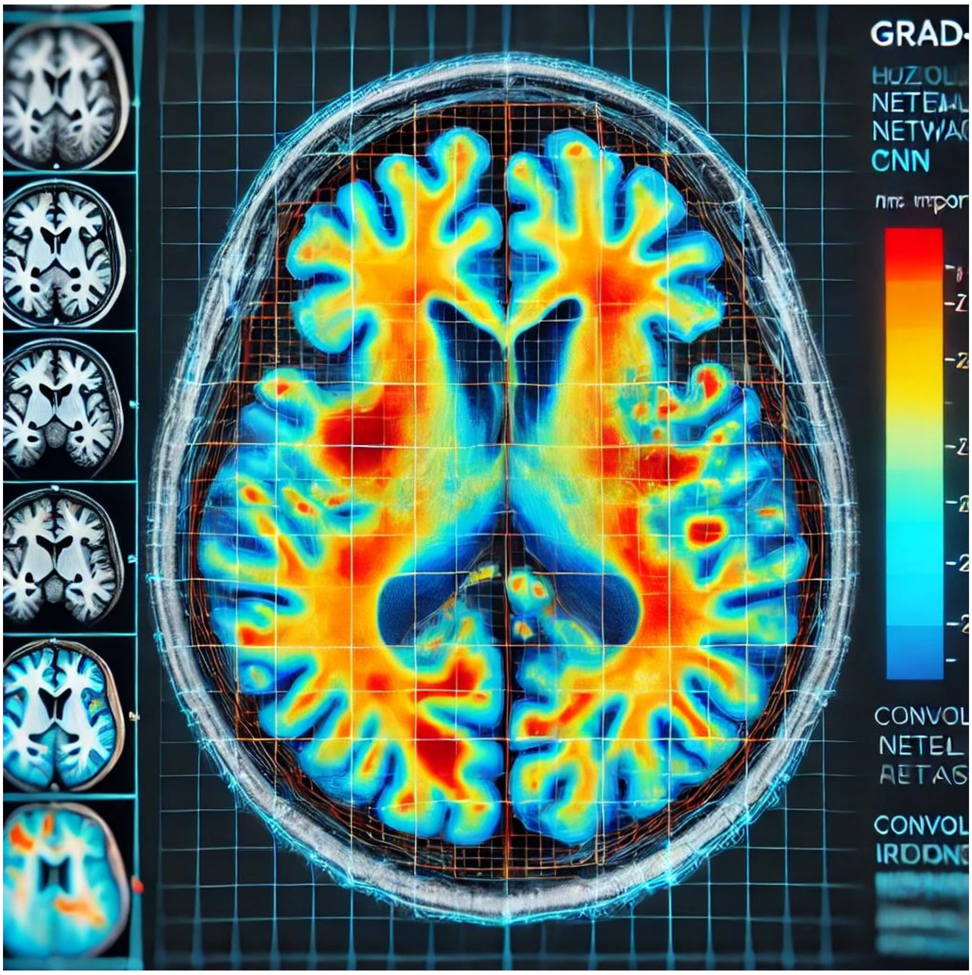
A 2024 study introduced the concept of XIAI to distinguish between explainable and interpretable models in healthcare NLP. The study found that attention mechanisms in NLP models can improve interpretability by highlighting relevant portions of clinical text that the model uses to make predictions. This approach enhances model transparency and can improve adoption in critical healthcare applications [53].

Example Use Case: In analyzing patient EHRs for predicting complications during surgery, an NLP model with attention-based XAI mechanisms can highlight the most relevant parts of the clinical text (e.g., past surgical history or medication use). This allows clinicians to understand how the model arrived at its predictions, increasing confidence in AI-supported decision-making.

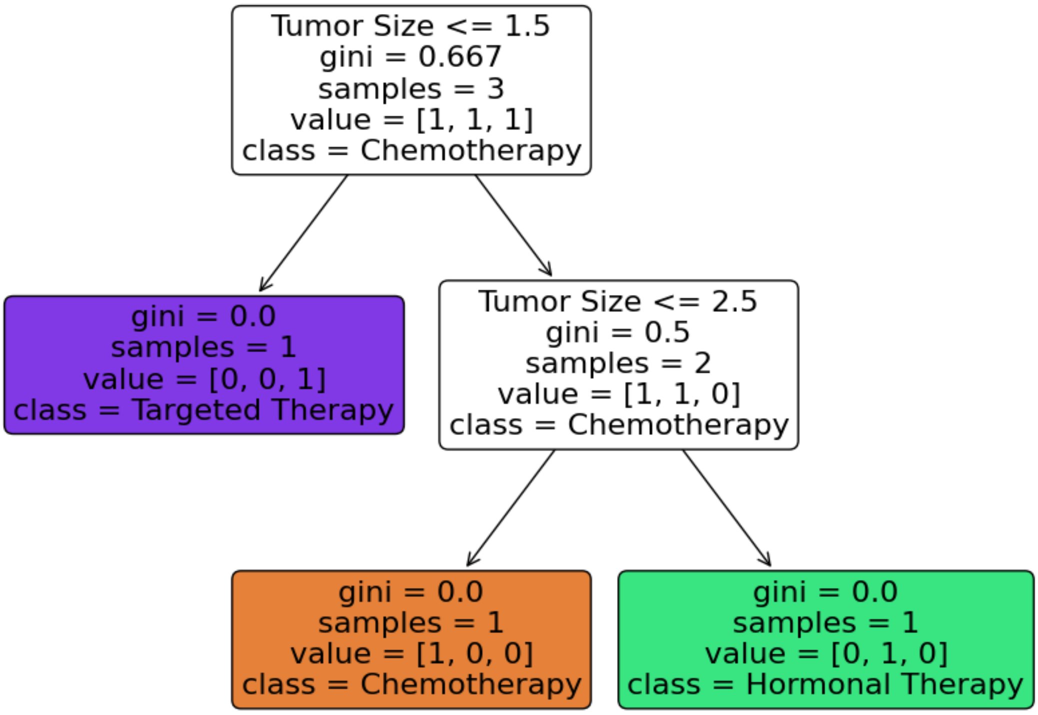
Table 3. Applications of XAI Techniques in Healthcare.

XAI Technique	Application	Purpose	Example Use Case	References
Grad-CAM	Medical Imaging (e.g., Alzheimer's, MRI)	Visualizes regions of medical images important for diagnosis.	Highlighting brain regions in MRI scans for Alzheimer's diagnosis using ensemble models.	[76,111]
Decision Trees with XAI	Oncology (e.g., Breast Cancer)	Provides transparent, personalized treatment recommendations.	Recommending personalized cancer therapies based on patient genetic markers and clinical data.	[109]
SHAP	Chronic Disease Risk Prediction	Identifies risk factors in predictive models.	Explaining factors like blood pressure and hospitalization history in heart failure readmission prediction.	[39]
Ensemble Transfer Learning & Vision Transformer	Disease Diagnosis	Enhances the accuracy and interpretability of complex models by combining multiple AI techniques.	Alzheimer’s disease diagnosis using hybrid models with higher accuracy and clearer interpretability.	[111]
XIAI (Explainable & Interpretable AI)	NLP in Healthcare	Improves model transparency in healthcare NLP tasks for decision-making.	Enhancing interpretability of large language models in personalized	[53]

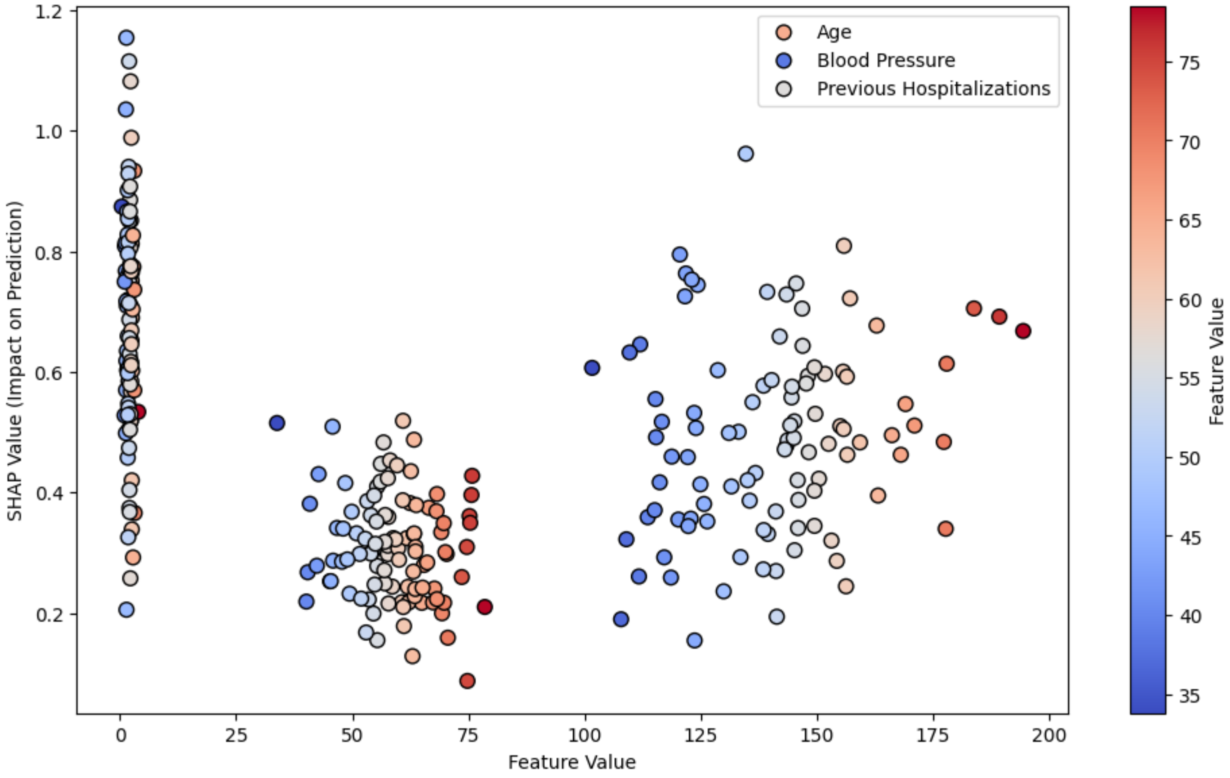
			medicine and medical task applications.	



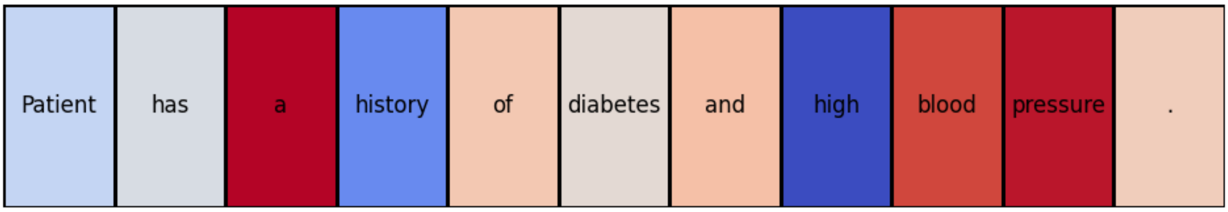
(A)


Tumor Size <= 1.5
gini = 0.667
samples = 3
value = [1, 1, 1]
class = Chemotherapy
gini = 0.0
samples = 1
value = [0, 0, 1]
class = Targeted Therapy
Tumor Size <= 2.5
gini = 0.5
samples = 2
value = [1, 1, 0]
class = Chemotherapy
gini = 0.0
samples = 1
value = [1, 0, 0]
class = Chemotherapy
gini = 0.0
samples = 1
value = [0, 1, 0]
class = Hormonal Therapy

(B)


Age
Blood Pressure
Previous Hospitalizations
SHAP Value (Impact on Prediction)
1.2
1.0
0.8
0.6
0.4
0.2
0
25
50
75
100
125
150
175
200
Feature Value
Feature Value
75
70
65
60
55
50
45
40
35

(C)



(D)

Figure 3. Illustrative Visualizations of XAI Techniques in Healthcare Diagnostics. (A) Grad-CAM Visualization for Alzheimer's Diagnosis; (B) Decision Tree for Breast Cancer Treatment Recommendation; (C) SHAP Values for Heart Failure Readmission Prediction; (D) Attention Mechanism Highlighting Relevant Words in EHR Analysis.

The heatmap in Figure 3A highlights areas of importance in the brain MRI, where red/orange regions indicate higher relevance in the CNN's decision-making process for diagnosing Alzheimer's. This helps clinicians confirm whether the AI model is focusing on clinically significant regions, improving the transparency of AI predictions. Figure 3B illustrates a Decision Tree for breast cancer treatment recommendation, with nodes representing key features such as tumor size, hormone receptor status, HER2 status, and patient age. Each branch leads to a treatment recommendation (chemotherapy, hormonal therapy, or targeted therapy), showing how the AI model arrives at personalized decisions based on clinical and genetic factors. This decision tree is a visual representation of how explainable AI models in oncology provide clear, interpretable pathways for treatment recommendations, making them transparent for clinicians. Figure 3C illustrates SHAP values for heart failure readmission prediction, showing the impact of three critical features: age, blood pressure, and previous hospitalizations. Each dot represents a patient, with the position on the x-axis indicating the feature values and the y-axis showing the SHAP values, which measure the contribution of each feature to the model's prediction. Positive SHAP values indicate a higher likelihood of readmission, while negative values suggest a lower likelihood. This visualization helps clinicians focus on key risk factors and offers transparency in AI predictions for better decision-making. Figure 3D illustrates the Attention Mechanism in NLP for EHR analysis, where the heatmap highlights relevant words in a patient's electronic health record (EHR). This mechanism helps the AI model focus on key terms such as "diabetes" and "blood pressure," which are crucial for making predictions about patient health. The attention scores provide transparency, allowing clinicians to understand which parts of the EHR the model considered most important.

5. Benefits and Limitations of XAI in Healthcare

XAI has emerged as a pivotal component in the development of AI systems for healthcare diagnostics, offering numerous advantages that improve trust, accountability, and clinical outcomes. However, despite its benefits, XAI also faces significant limitations, particularly when applied to complex medical environments. Below is a detailed examination of both the benefits and limitations of XAI in healthcare.

5.1. Benefits of XAI in Healthcare

5.1.1. Enhanced Trust and Transparency

One of the foremost benefits of XAI in healthcare is its ability to enhance trust and transparency in AI systems. Traditional AI models, such as deep neural networks, often function as "black boxes," making it difficult for clinicians to understand how a diagnosis or treatment recommendation is made. XAI techniques, such as Grad-CAM in medical imaging or SHAP in chronic disease prediction, allow clinicians to visualize or quantify how the model arrived at a particular decision. This interpretability builds trust between healthcare professionals and AI systems, ensuring that decisions are understandable and aligned with clinical knowledge.

For example, in an AI system predicting cancer treatment pathways, a decision tree model enhanced by XAI can clearly display the factors contributing to the recommendation of chemotherapy versus hormonal therapy. Clinicians can see the influence of tumor size, genetic markers, and hormone receptor status, leading to more informed and confident clinical decisions [109].

5.1.2. Improved Regulatory Compliance and Ethical AI

In healthcare, regulatory standards, such as those set by the U.S. Food and Drug Administration (FDA) or European Medicines Agency (EMA), require AI systems to provide interpretable decisions to ensure patient safety and accountability. XAI techniques help meet these regulatory demands by providing insights into how and why certain decisions were made, enabling greater oversight and validation. Explainable models can justify the reasoning behind high-risk decisions such as surgical procedures, drug prescriptions, or disease prognoses, reducing liability risks for healthcare providers and aligning AI systems with ethical standards [111].

5.1.3. Personalized and Precise Patient Care

XAI allows for more personalized and precise patient care by offering detailed insights into individual-level predictions. In personalized medicine, AI models can tailor treatment plans based on patient-specific clinical and genetic data. For example, in precision oncology, XAI-driven models can highlight the genetic markers that led to specific treatment recommendations, allowing clinicians to tailor therapy to the patient's unique genetic profile. This level of personalization is essential for improving treatment outcomes and minimizing adverse reactions [39].

5.1.4. Facilitates Human-AI Collaboration

XAI promotes better collaboration between AI systems and clinicians by making AI outputs more interpretable and actionable. By providing clear rationales for predictions, XAI enables healthcare professionals to integrate AI insights into their own expertise and decision-making processes. For instance, in radiology, Grad-CAM heatmaps help radiologists see which regions of an MRI the AI model has focused on, allowing them to verify the diagnosis or investigate further [76]. This symbiosis between AI and human expertise leads to more robust diagnostic outcomes.

5.2. Limitations of XAI in Healthcare

5.2.1. Complexity of Interpretability

While XAI aims to make AI models more interpretable, the interpretability itself can sometimes be complex and difficult for non-experts to understand. Techniques like SHAP or LIME, while powerful, produce explanations that may not always be intuitively clear to clinicians without deep technical knowledge. For example, while SHAP values can provide a feature-based explanation for a prediction, understanding the underlying mechanics of Shapley values and how they relate to cooperative game theory can be challenging in a fast-paced clinical setting [75]. Therefore, XAI systems must balance complexity with usability to ensure that the explanations provided are genuinely useful in practice.

5.2.2. Scalability Issues in Large Models

One significant limitation of XAI techniques is their scalability, particularly in large and complex AI models like deep learning networks with millions of parameters. Techniques like Grad-CAM and SHAP, although highly effective for small- to medium-sized models, can become computationally expensive and slow when applied to large-scale networks used in medical diagnostics. This can limit

doi:10.20944/preprints202409.1974.v

the practical deployment of XAI in real-time clinical environments, where speed and accuracy are crucial [39].

5.2.3. Limitations in Explaining Certain AI Models

Not all AI models lend themselves easily to explanation through XAI techniques. For instance, while models like decision trees or linear regression are inherently interpretable, more complex models such as deep reinforcement learning or RNNs used for time-series medical data often remain difficult to explain using current XAI methods. This presents a challenge in certain healthcare applications, such as long-term patient monitoring or predictive modeling based on temporal data, where clinicians may struggle to interpret and validate the model's decisions.

5.2.4. Risk of Over-Simplification

In efforts to make AI models more interpretable, there is a risk that XAI techniques may oversimplify the underlying logic, potentially leading to inaccurate or misleading conclusions. For example, in some cases, post-hoc explanation methods like LIME may provide locally accurate explanations that do not fully reflect the global behavior of the model. This could lead clinicians to make decisions based on incomplete or misleading insights, which could compromise patient care [112].

5.2.5. Ethical and Bias Concerns

XAI does not fully eliminate ethical concerns around AI systems, particularly regarding bias. While XAI can make the decision-making process more transparent, it does not inherently address the issue of biased data or biased decision-making within the models themselves. If the underlying AI system is trained on biased datasets, even the explanations provided by XAI may reflect and perpetuate these biases, potentially leading to discriminatory outcomes in healthcare, particularly for underrepresented groups [53].

In summary, the integration of XAI in healthcare brings a multitude of benefits, from increasing trust and transparency to supporting personalized patient care and ensuring regulatory compliance. However, the limitations of XAI, including interpretability challenges, scalability issues, and ethical concerns, need to be carefully considered in its adoption. As AI continues to evolve, so too must the techniques for making these systems explainable, ensuring that they are both powerful and practical for real-world clinical applications.

Table 4 summarizes the primary benefits and limitations of applying XAI in healthcare diagnostics. It contrasts how XAI improves trust, transparency, regulatory compliance, and personalized care, while also addressing challenges such as scalability, interpretability complexity, and potential biases.

Table 4. Summary of Benefits and Limitations of XAI in Healthcare.

Aspect	Benefits	Limitations	References
Trust and Transparency	Enhances clinician trust with interpretable decisions.	Interpretability complexity can hinder understanding for non-experts.	[75,76]
Regulatory Compliance	Supports regulatory requirements for safe and accountable AI.	Not all AI models can be easily explained (e.g., RNNs).	[111]
Personalized Care	Allows more precise and tailored treatment	Over-simplification in post-hoc methods	[39,109]

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 September 2024 doi:10.20944/preprints202409.1974.v

	plans for individual patients.	may mislead clinicians.	
Human-AI Collaboration	Facilitates better collaboration between AI systems and clinicians.	Scalability issues in larger models affect real-time performance.	[53]
Ethical Considerations	Helps address bias and fairness by providing insights into model decisions.	XAI may still perpetuate biases present in the training data.	[53,112]

Key Points:

- This table clearly outlines both the advantages and challenges associated with using XAI in healthcare, making it easier for readers to compare both sides.
- The benefits cover trust-building, regulatory support, and personalized care, while the limitations focus on interpretability, scalability, and ethical concerns.

Figure 4A workflow diagram illustrates how XAI integrates into healthcare diagnostics. It begins with patient data input (e.g., medical images or EHR), followed by AI model processing, the application of XAI techniques (e.g., Grad-CAM, SHAP), the generation of explainable outputs (heatmaps, feature importance scores), and concludes with human decision-making, where clinicians review the explainable outputs to make final diagnoses or treatment decisions.

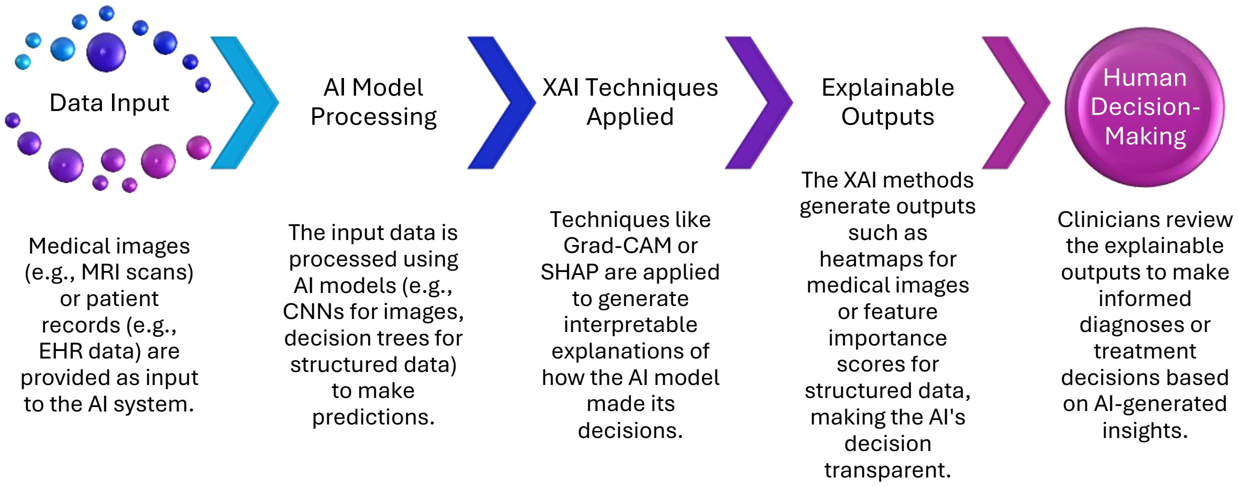
Key Points:

- The workflow shows the step-by-step process in which XAI techniques are embedded in the healthcare decision-making pipeline.
- It emphasizes how XAI provides interpretable insights that allow clinicians to verify AI-generated recommendations, improving trust and reliability in AI-based diagnostics.

Figure 4B graph shows the relationship between model complexity and the scalability and interpretability of XAI techniques. As AI models become more complex (e.g., transitioning from decision trees to CNNs or transformers), the scalability and interpretability of XAI techniques decrease. The graph highlights the performance trade-offs between maintaining explainability and handling larger, more complex models.

Key Points:

- Simpler models (e.g., decision trees) are easier to interpret but may not scale well for large datasets or complex tasks.
- More complex models (e.g., CNNs, Transformers) offer higher accuracy but are harder to interpret and require more computational resources, limiting their scalability.
- The graph helps visualize how XAI methods like SHAP and Grad-CAM perform in relation to model complexity and provides insight into the trade-offs between interpretability and computational efficiency.



Medical images (e.g., MRI scans) or patient records (e.g., EHR data) are provided as input to the AI system.

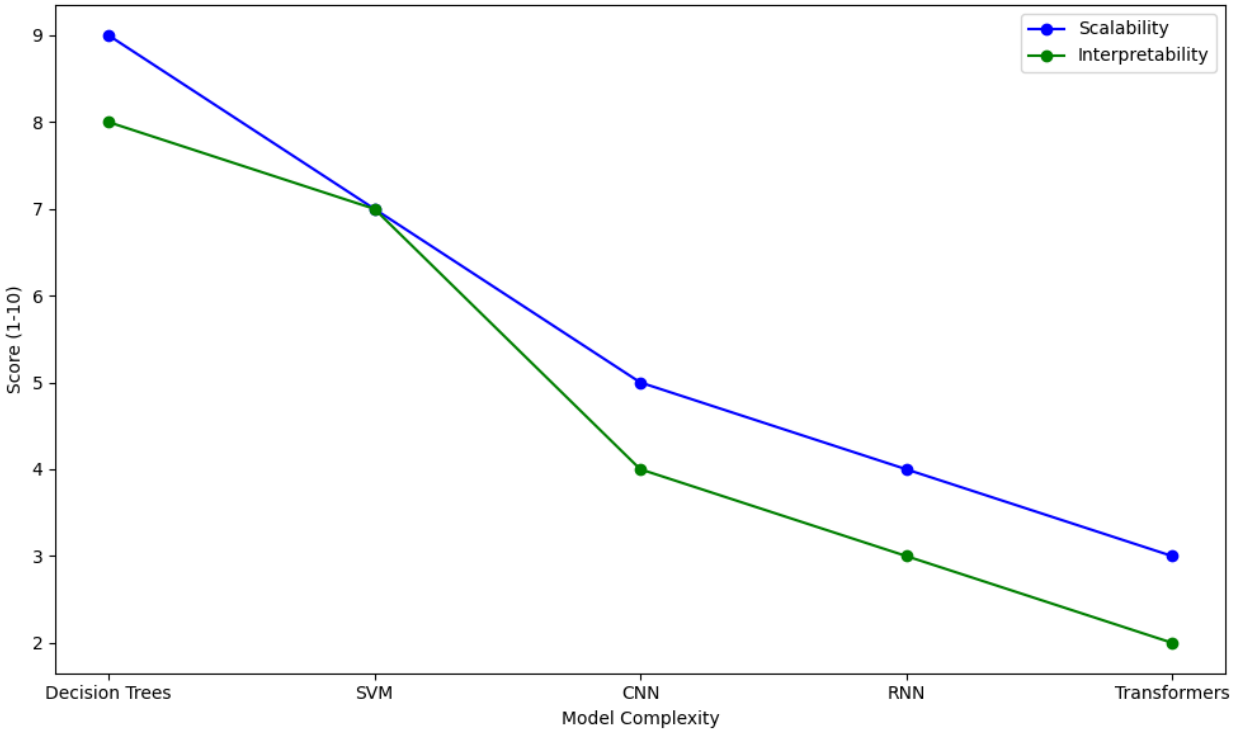
The input data is processed using AI models (e.g., CNNs for images, decision trees for structured data) to make predictions.

Techniques like Grad-CAM or SHAP are applied to generate interpretable explanations of how the AI model made its decisions.

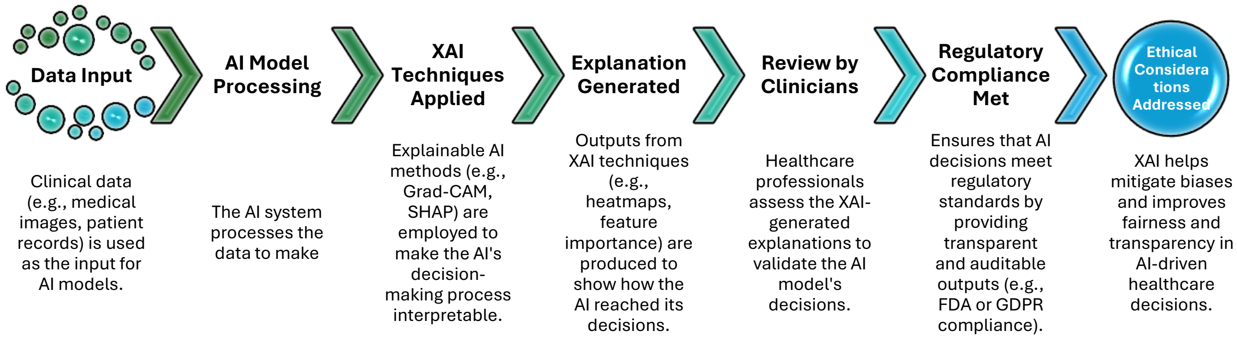
The XAI methods generate outputs such as heatmaps for medical images or feature importance scores for structured data, making the AI's decision transparent.

Clinicians review the explainable outputs to make informed diagnoses or treatment decisions based on AI-generated insights.

(A)



(B)



Clinical data (e.g., medical images, patient records) is used as the input for AI models.

The AI system processes the data to make

Explainable AI methods (e.g., Grad-CAM, SHAP) are employed to make the AI's decision-making process interpretable.

Outputs from XAI techniques (e.g., heatmaps, feature importance) are produced to show how the AI reached its decisions.

Healthcare professionals assess the XAI-generated explanations to validate the AI model's decisions.

Ensures that AI decisions meet regulatory standards by providing transparent and auditable outputs (e.g., FDA or GDPR compliance).

XAI helps mitigate biases and improves fairness and transparency in AI-driven healthcare decisions.

doi:10.20944/preprints202409.1974.v

(C)

Figure 4. Illustrative Visuals for Benefits and Limitations of XAI in Healthcare. (A) Workflow of XAI in Healthcare Diagnostics; (B) Scalability of XAI Techniques vs Model Complexity; (C) Flowchart XAI Impact on Regulatory Compliance and Ethical Considerations.

Figure 4C flowchart illustrates the role of XAI in ensuring regulatory compliance and addressing ethical concerns such as bias and transparency in healthcare. The diagram shows how data is processed through AI models, explained using XAI techniques, and reviewed by clinicians, ultimately achieving regulatory compliance (e.g., FDA, GDPR) and addressing ethical considerations (e.g., bias detection).

Key Points:

- XAI techniques play a vital role in aligning AI models with regulatory frameworks, ensuring transparency and auditability.
- The flowchart highlights how ethical concerns, such as fairness and bias in AI decisions, can be mitigated by applying XAI methods, ultimately leading to safer and more accountable healthcare systems.

General Observations for All Visual Elements:

- ***Context:*** These visuals collectively explain how XAI contributes to improving healthcare diagnostics while simultaneously addressing its limitations, including regulatory and ethical challenges.
- ***Application:*** They highlight how XAI fits into healthcare workflows, balancing the benefits of interpretability with the practical challenges of scalability and regulatory requirements.
- ***Importance:*** The visuals emphasize how XAI techniques such as Grad-CAM, SHAP, and LIME help bridge the gap between AI model complexity and the need for transparent, reliable decision-making in clinical environments.

6. Ethical and Regulatory Considerations in XAI for Healthcare

As AI continues to be adopted in healthcare, ensuring that AI systems are ethical, transparent, and compliant with regulations has become a critical concern. XAI plays a crucial role in addressing both ethical and regulatory challenges by making the decision-making processes of AI models more understandable. Below is an exploration of the ethical implications and regulatory requirements for XAI in healthcare.

6.1. Ethical Considerations in XAI

6.1.1. Addressing Bias and Fairness

One of the most pressing ethical concerns in healthcare AI is the potential for biased or unfair decision-making. If an AI model is trained on data that underrepresents certain populations or is skewed toward specific demographic groups, it may produce biased results. For instance, an AI system trained primarily on data from one ethnic group may not perform as well on patients from other ethnicities. XAI techniques, such as SHAP or LIME, provide visibility into which features the AI model considers important for making predictions, helping to identify and correct for bias.

In recent studies, XAI has been used to expose bias in models used for predicting disease risks in minority populations. By making the decision-making process transparent, XAI allows researchers and clinicians to scrutinize whether certain features (e.g., race, gender) disproportionately affect the AI's decisions and adjust the model accordingly [53].

6.1.2. Ensuring Transparency

Transparency is a key ethical concern in healthcare AI. Patients and clinicians must understand how AI systems arrive at their decisions, especially when these decisions have life-or-death consequences, such as diagnosing cancer or recommending surgical interventions. XAI techniques provide the necessary tools to increase transparency by explaining the reasoning behind AI-driven decisions. In turn, this fosters trust between clinicians, patients, and AI systems.

Grad-CAM, for example, can show exactly which parts of a medical image a model focuses on when diagnosing a condition like Alzheimer's. This helps radiologists verify whether the AI's focus aligns with known clinical indicators, improving the reliability and ethical integrity of AI-assisted diagnostics [76].

6.1.3. Accountability and Trust

In healthcare, where incorrect diagnoses or treatment recommendations can lead to severe consequences, accountability is paramount. XAI enhances accountability by allowing clinicians to understand and question the AI's decision-making process, ensuring that decisions are grounded in clinical reasoning. This is particularly important for healthcare providers, as they are ultimately responsible for patient outcomes. If an AI system's decisions are opaque, clinicians may hesitate to trust it, which can hinder adoption. XAI helps bridge this gap by making AI systems more trustworthy.

A key application is in personalized medicine, where clinicians rely on AI to make recommendations for tailored treatments. By using XAI methods like decision trees or SHAP, clinicians can understand which factors influenced the treatment suggestion, thereby making the AI system more accountable and improving its integration into healthcare workflows [109].

6.2. Regulatory Considerations in XAI

6.2.1. Meeting Regulatory Standards

AI systems in healthcare must comply with various regulatory standards to ensure safety, accuracy, and transparency. Regulatory bodies such as the U.S. Food and Drug Administration (FDA) and the European Medicines Agency (EMA) require AI models to be transparent and auditable. XAI techniques are crucial for meeting these standards by providing interpretable outputs that can be scrutinized by regulators.

For example, the FDA's AI/ML-based Software as a Medical Device (SaMD) Action Plan emphasizes the need for transparency and continuous learning in AI systems. XAI methods can provide the necessary insights into how an AI system makes decisions, allowing regulators to assess the model's safety and reliability before approving it for clinical use [111].

6.2.2. Audibility and Validation

XAI not only improves transparency but also enhances the audibility of AI systems in healthcare. Regulators require that AI systems be auditable, meaning that their decisions can be tracked and validated. This is especially important in sensitive areas like diagnostics or treatment recommendations, where an audit trail ensures that decisions are well-founded and based on sound medical principles.

By using XAI, healthcare providers can maintain an audit trail of the AI's decision-making process. For instance, if an AI model recommends a particular treatment, XAI can highlight the key factors that influenced the decision. This audit trail can be reviewed by both regulators and clinicians to ensure the AI system is making decisions aligned with regulatory standards.

6.2.3. Data Privacy and Security

Data privacy is a key regulatory concern, particularly with the advent of laws like General Data Protection Regulation (GDPR) in Europe and the Health Insurance Portability and Accountability Act (HIPAA) in the United States. These regulations require that patient data be handled securely and that any decisions made using AI systems are explainable, especially when personal data is involved. XAI techniques ensure that AI models comply with these regulations by providing clear explanations about how sensitive data is used in decision-making.

XAI helps balance the need for data-driven insights with the requirement to protect patient privacy. For example, if a model uses patient demographics to make a prediction, XAI can ensure that this data is used appropriately and that the model's decision-making process complies with relevant privacy laws [39].

6.3. The Future of XAI in Ethical and Regulatory Frameworks

As AI continues to evolve, so too must the ethical and regulatory frameworks surrounding it. In the future, regulatory bodies may require not only explainable AI models but also continuous monitoring of AI systems to ensure they remain transparent and unbiased over time. As AI systems learn and evolve with new data, it will be critical to maintain their explainability to ensure compliance with ethical and regulatory standards.

Future frameworks may also prioritize responsible AI principles, which emphasize fairness, transparency, and accountability. XAI will play an integral role in ensuring that AI systems in healthcare align with these principles, ensuring that they contribute to better patient outcomes while safeguarding ethical integrity.

In summary, XAI serves as a vital tool in addressing the ethical and regulatory challenges posed by AI in healthcare. By improving transparency, mitigating bias, and providing accountability, XAI helps build trust in AI systems while ensuring they comply with regulatory requirements. As AI continues to integrate into healthcare, XAI will be key in shaping ethical AI practices and ensuring that AI systems are both effective and responsible.

Table 5. Ethical Considerations in XAI for Healthcare.

Ethical Aspect	**Description**	**XAI Technique Addressing the Concern**	**References**
Bias and Fairness	Identifies biased or unfair decision-making in AI models due to imbalanced data or flawed feature selection.	SHAP for feature attribution, LIME for local explanations.	[53]
Transparency	Ensures that clinicians and patients can understand how AI systems arrive at their decisions.	Grad-CAM for visual explanations in medical imaging.	[76]
Accountability	Improves trust by providing clinicians with the ability to trace and verify AI decisions.	Decision Trees, SHAP for global interpretability.	[109]
Data Privacy	Ensures that patient data is handled securely and used appropriately in AI models.	Data usage transparency through XAI outputs (SHAP, LIME).	[39]

Table 5 summarizes the primary ethical concerns in applying AI in healthcare, such as bias, transparency, accountability, and data privacy. It highlights how specific XAI techniques (like SHAP and Grad-CAM) address these concerns. For example, SHAP helps identify biased feature attribution in patient outcomes, while Grad-CAM enhances transparency in medical image analysis by providing visual explanations of AI decisions. This table serves as a quick reference for understanding how XAI mitigates ethical risks and improves trust in AI-driven decisions.

Key Points:

- Ethical challenges are directly linked to the decision-making process of AI models.
- XAI techniques provide transparency, fairness, and auditability in healthcare applications.

Figure 5A flowchart outlines the critical steps in ensuring that XAI models meet healthcare regulatory standards. The workflow begins with input data (e.g., medical images or patient records), processed through AI models (CNNs or decision trees). XAI techniques (e.g., Grad-CAM, SHAP) are applied to generate explainable outputs, such as heatmaps or feature importance scores, which are reviewed by clinicians. The final stage includes validation and auditing by regulators, ensuring compliance with frameworks such as FDA, GDPR, and HIPAA.

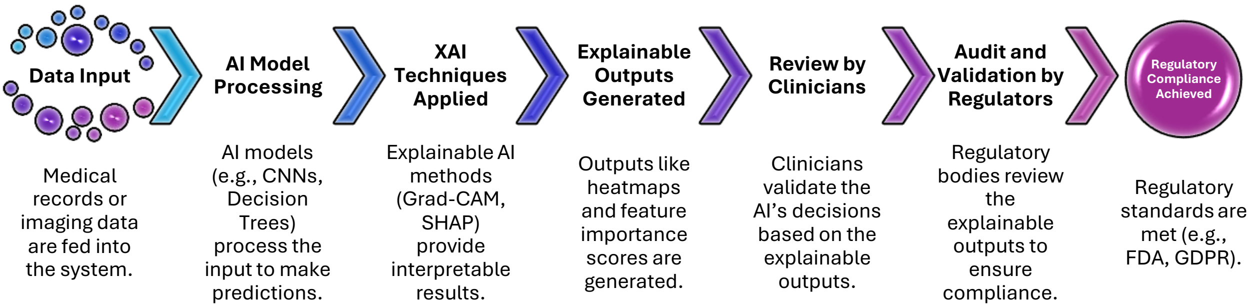
Key Points:

- XAI is instrumental in achieving transparency and explainability for regulatory audits.
- Clinicians and regulators can scrutinize the model's decisions, ensuring that they are aligned with clinical and legal standards.

Figure 5B line graph shows the growing importance of XAI in healthcare from 2018 to 2024. The upward trend reflects the increasing demand for explainability as AI systems become more prevalent in medical practice. Regulatory bodies such as the FDA and EMA are progressively requiring AI systems to offer interpretability to meet safety, fairness, and transparency criteria. The graph highlights that, over time, XAI is playing a more significant role in aligning AI technologies with ethical and regulatory standards.

Key Points:

- The rising curve represents the heightened focus on XAI by regulators and healthcare institutions.
- XAI is critical in achieving regulatory approval for AI-based healthcare applications as explainability becomes a primary concern.



Data Input

Medical records or imaging data are fed into the system.

AI Model Processing

AI models (e.g., CNNs, Decision Trees) process the input to make predictions.

XAI Techniques Applied

Explainable AI methods (Grad-CAM, SHAP) provide interpretable results.

Explainable Outputs Generated

Outputs like heatmaps and feature importance scores are generated.

Review by Clinicians

Clinicians validate the AI's decisions based on the explainable outputs.

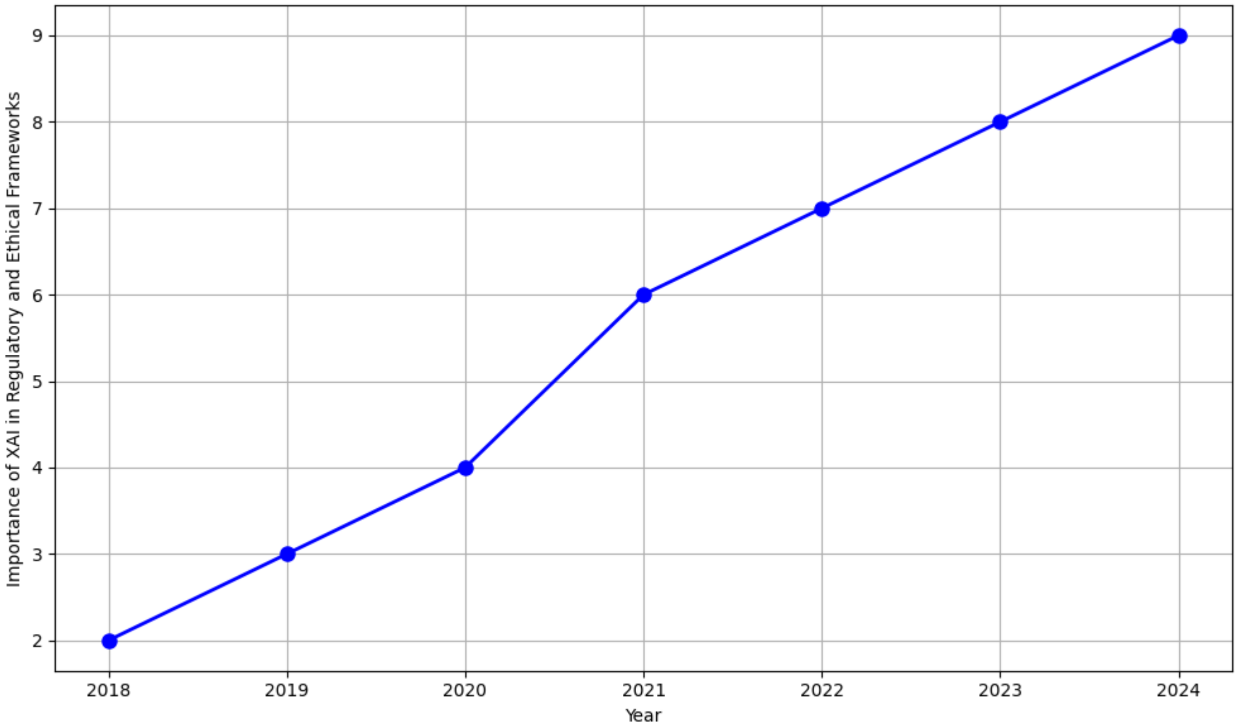
Audit and Validation by Regulators

Regulatory bodies review the explainable outputs to ensure compliance.

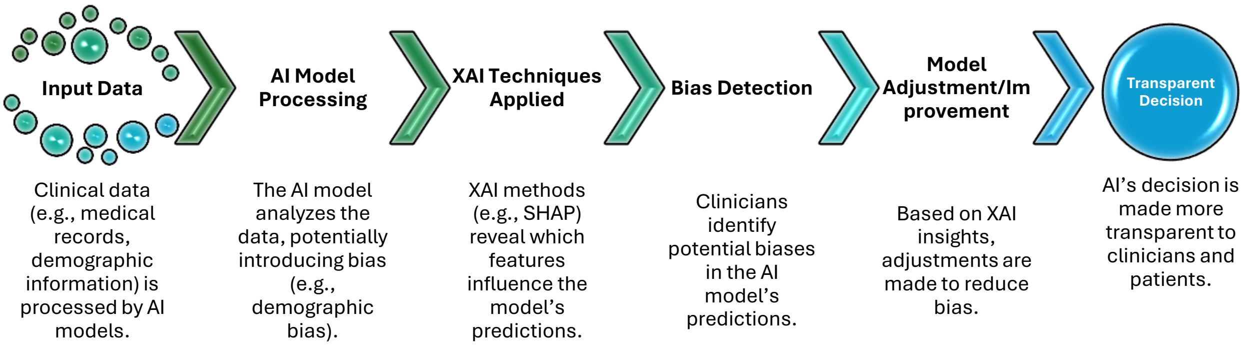
Regulatory Compliance Achieved

Regulatory standards are met (e.g., FDA, GDPR).

(A)



(B)



Clinical data (e.g., medical records, demographic information) is processed by AI models.

The AI model analyzes the data, potentially introducing bias (e.g., demographic bias).

XAI methods (e.g., SHAP) reveal which features influence the model's predictions.

Clinicians identify potential biases in the AI model's predictions.

Based on XAI insights, adjustments are made to reduce bias.

AI's decision is made more transparent to clinicians and patients.

(C)

Figure 5. Illustrative Visuals for Ethical and Regulatory Considerations in XAI for Healthcare. (A) Workflow of XAI's Role in Regulatory Compliance; (B) Impact of XAI on Ethical and Regulatory Frameworks (2018-2024); (C) XAI for Bias Mitigation and Transparency in Decision-Making.

Figure 5C diagram illustrates how XAI techniques, like SHAP, help mitigate bias and improve transparency in healthcare AI systems. The workflow shows how data, when processed by an AI model, may introduce bias. XAI techniques are then applied to uncover the factors that influence the AI's decision, helping clinicians detect and address any biases. By adjusting the model based on XAI feedback, clinicians can ensure fair and transparent decisions for patient care.

Key Points:

- XAI helps expose biases that may arise in AI-driven predictions (e.g., racial or gender bias).
- The diagram showcases how XAI can be integrated into AI workflows to ensure that decision-making is fair and transparent.

General Observations for All Visual Elements:

- ***Context:*** These visual elements demonstrate the dual role of XAI in ensuring compliance with both ethical standards (like fairness, transparency, and accountability) and regulatory requirements (such as those imposed by the FDA, GDPR, and HIPAA).
- ***Application:*** They highlight that XAI techniques are essential in providing explainability, which is crucial for the widespread adoption of AI systems in clinical environments.
- ***Importance:*** The visuals emphasize the growing reliance on XAI to ensure that AI models in healthcare are interpretable, compliant, and ethically aligned with patient safety and care.

7. Discussion

XAI has been a game-changer in healthcare diagnostics, offering insights into the often opaque decision-making processes of advanced AI models. While XAI techniques like SHAP, Grad-CAM, and LIME have bridged the gap between AI model outputs and clinician understanding, the implementation of XAI in real-world healthcare systems still faces significant challenges. This discussion addresses the core benefits and limitations of XAI, focusing on trust, accountability, bias mitigation, scalability, and regulatory compliance.

7.1. Balancing Accuracy and Interpretability

One of the most persistent challenges in applying XAI in healthcare is the trade-off between the accuracy of complex models (like deep learning models) and their interpretability. Advanced AI models, such as CNNs and transformers, are known for their high accuracy in tasks like medical imaging analysis, but they often function as "black boxes," making their decisions hard to explain and understand [76]. On the other hand, simpler models like decision trees are more interpretable but may not offer the same level of accuracy for complex tasks [113]. XAI techniques such as SHAP have been introduced to provide local explanations for individual predictions, enabling clinicians to see which features most influenced the model's decision [114].

Despite these advancements, current XAI methods are computationally expensive, which limits their application in real-time diagnostics where speed is critical [115]. Moreover, scalability remains a challenge, as models increase in complexity, and ensuring that XAI can function efficiently in time-sensitive clinical settings requires further research [116].

7.2. Trust and Accountability in AI-Driven Healthcare

Building trust between clinicians and AI systems is crucial for the widespread adoption of AI in healthcare. Studies show that clinicians are more likely to use AI systems if they can understand and verify the reasoning behind the decisions. Tools like Grad-CAM have been particularly helpful in medical imaging, allowing clinicians to visualize which parts of an MRI or CT scan the model focused on for diagnosis [76]. This has improved the transparency of AI in radiology and increased clinician confidence in AI-based diagnostic tools [117].

However, interpretability alone does not ensure accountability. AI systems are not immune to errors, and when these errors occur, particularly in critical areas like diagnosis or treatment recommendations, clinicians need to trace the logic behind the decision to determine whether it was valid [112]. XAI techniques allow this, but further work is needed to define clear accountability frameworks, particularly in cases where AI-driven decisions lead to adverse outcomes [2].

7.3. Ethical Considerations and Bias Mitigation

AI models trained on biased datasets can perpetuate and even exacerbate existing health disparities, leading to unequal treatment of patients. This is especially concerning in healthcare, where underrepresented populations may be at risk of receiving suboptimal care due to biased AI predictions [118]. XAI techniques offer a way to identify and mitigate bias by providing transparency in model predictions. For instance, SHAP can highlight whether certain demographic features (such as race or gender) disproportionately influence the model's decision-making process [39]. This allows for adjustments to be made to the model or the data to ensure fairer outcomes.

However, XAI cannot eliminate all biases inherent in data. The presence of biases in medical datasets remains a significant issue that XAI alone cannot solve. Ethical guidelines and rigorous data collection protocols are essential to ensure that AI systems do not reinforce discriminatory practices [119]. Future research should focus on the development of ethical AI frameworks that prioritize fairness and accountability while still leveraging the advantages of XAI techniques [120].

7.4. Regulatory Compliance and Legal Implications

The increasing use of AI in healthcare has attracted attention from regulatory bodies, such as the FDA and the European Medicines Agency (EMA), which emphasize the need for transparency and explainability in AI systems [111]. XAI plays a critical role in ensuring that AI models meet these regulatory requirements by providing interpretable outputs that can be scrutinized for safety and effectiveness [121]. Legal frameworks are also beginning to address the role of AI in clinical decision-making, particularly in defining who is accountable when an AI system contributes to a medical error [122].

Additionally, compliance with privacy regulations, such as GDPR and HIPAA, is crucial when using patient data in AI systems [123]. XAI systems must not only explain their decisions but also ensure that sensitive patient data is protected and used in compliance with legal standards. As regulatory frameworks evolve, XAI will need to adapt to provide explanations that meet these rigorous standards while maintaining patient confidentiality [116].

7.5. The Future of XAI in Healthcare

The future of XAI in healthcare lies in the development of hybrid models that combine the interpretability of simpler models with the accuracy of more complex deep learning systems [113]. Explainable-by-design AI systems, which inherently incorporate transparency, may reduce the reliance on post-hoc explanation techniques like SHAP and Grad-CAM [124]. These models are more likely to be integrated into real-time clinical workflows, where speed and accuracy are paramount [125].

Moreover, there is a growing need for intuitive user interfaces that allow clinicians to interact with XAI models easily, facilitating the integration of AI into day-to-day clinical decision-making [2]. As the field of AI continues to evolve, the development of scalable, efficient XAI methods will be key to ensuring that AI is trusted, reliable, and ethically sound in healthcare.

In summary, this discussion highlights the critical role XAI plays in healthcare, improving transparency, trust, and accountability while addressing bias and regulatory challenges. However, much work remains to be done, particularly in balancing accuracy with interpretability, mitigating bias, and developing scalable solutions for real-time clinical use. As AI continues to transform healthcare, XAI will be indispensable in ensuring that these technologies are not only powerful but also ethically aligned with patient needs.

8. Conclusion

XAI is transforming healthcare diagnostics by addressing one of the most pressing challenges in artificial intelligence—understanding and interpreting the complex decision-making processes of AI models. By making AI models more transparent and interpretable, XAI enhances trust, accountability, and collaboration between clinicians and AI systems. However, despite these promising advancements, significant challenges remain, particularly around the trade-off between accuracy and interpretability, scalability of XAI techniques, bias mitigation, and regulatory compliance.

One of the key benefits of XAI is its ability to improve clinician trust in AI systems by providing interpretable outputs that explain how and why a model arrived at a particular diagnosis or recommendation. Techniques such as Grad-CAM, SHAP, and LIME have enabled clinicians to visualize and understand AI decisions, fostering greater confidence in AI-assisted diagnostics. This is particularly important in fields such as radiology and oncology, where AI systems are increasingly relied upon to detect diseases early and recommend personalized treatments. Nevertheless, the complexity of interpretability, especially in deep learning models like CNNs and transformers, remains a critical challenge.

Another advantage of XAI is its potential to enhance regulatory compliance. As AI becomes more integral to clinical workflows, regulatory bodies such as the FDA and EMA are emphasizing the importance of explainability in AI models. XAI techniques help ensure that AI systems can meet these regulatory requirements, providing transparency and auditability that are essential for safe and ethical deployment in healthcare. However, scalability continues to pose a problem, as more complex models often require extensive computational resources to generate interpretable explanations, limiting their real-time application in clinical settings.

Ethical considerations, particularly around bias and fairness, also play a central role in the XAI discussion. While XAI techniques provide a pathway for identifying and mitigating bias, they cannot fully eliminate biases embedded in datasets. As AI models are only as good as the data they are trained on, biased datasets will inevitably lead to biased predictions. This highlights the need for more rigorous data collection processes and continuous bias audits throughout the AI lifecycle. Regulatory frameworks will need to evolve to address these issues, ensuring that AI systems are not only effective but also fair and equitable for all patient groups.

Looking to the future, the development of hybrid models that combine the high accuracy of deep learning with the interpretability of simpler models may offer a promising solution to some of these challenges. Moreover, advancements in explainable-by-design AI models, which are inherently interpretable without requiring post-hoc explanation techniques, could further enhance the applicability of AI in healthcare. As AI technology continues to evolve, XAI will remain an essential component, ensuring that AI systems are transparent, accountable, and aligned with the ethical standards required in healthcare.

In conclusion, while XAI has made significant strides in making AI-driven healthcare diagnostics more interpretable and trustworthy, it is clear that ongoing research is needed to address the remaining challenges. Future efforts should focus on improving the scalability of XAI techniques, reducing computational costs, and enhancing bias detection. At the same time, collaboration between AI developers, clinicians, and regulators will be crucial to ensuring that XAI systems are seamlessly integrated into healthcare, ultimately improving patient outcomes and ensuring the ethical deployment of AI technologies.

Funding: This research received no external funding.

Compliance with Ethical Standards: This article does not involve any studies conducted by the authors that included human participants.

Acknowledgments: The completion of this research work was made possible through the collaborative efforts and dedication of a multidisciplinary team. We extend our sincere appreciation to each member for their invaluable contributions.

Conflicts of Interest: The authors declare no conflicts of in-terest.

References

1. Litjens, G., et al., *A survey on deep learning in medical image analysis.* Medical image analysis, 2017. **42**: p. 60-88.
2. Topol, E.J., *High-performance medicine: the convergence of human and artificial intelligence.* Nature medicine, 2019. **25**(1): p. 44-56.
3. Zech, J.R., et al., *Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study.* PLoS medicine, 2018. **15**(11): p. e1002683.
4. Samek, W., et al., *Explainable AI: interpreting, explaining and visualizing deep learning*. Vol. 11700. 2019: Springer Nature.
5. Abbasian Ardakani, A., et al., *Interpretation of Artificial Intelligence Models in Healthcare.* JOURNAL OF ULTRASOUND IN MEDICINE, 2024.

6. Chowdhury, N., *Intelligent systems for healthcare diagnostics and treatment.* World Journal of Advanced Research and Reviews, 2024. **23**(1): p. 007-015.
7. Singh, S., et al., *The Role of Artificial Intelligence in Treatment and Diagnosis in Healthcare.* Journal for Research in Applied Sciences and Biotechnology, 2024. **3**(4): p. 5-13.
8. Khinvasara, T., K.M. Cuthrell, and N. Tzenios, *Harnessing Artificial Intelligence in Healthcare Analytics: From Diagnosis to Treatment Optimization.* Asian Journal of Medicine and Health, 2024. **22**(8): p. 15-31.
9. Shaheen, M.Y., *Applications of Artificial Intelligence (AI) in healthcare: A review.* ScienceOpen Preprints, 2021.
10. Murali^1, N. and N. Sivakumaran, *Artificial intelligence in healthcare–a review.* 2018.
11. Bhagwan, J., et al., *An enhance CNN model for brain tumor detection from MRI images.* Journal of Electrical Systems, 2024. **20**(3s): p. 1072-1081.
12. Desai, C.K., et al., *Diagnosis of Medical Images Using Convolutional Neural Networks.* Journal of Electrical Systems, 2024. **20**(6s): p. 2371-2376.
13. Ebrahimi, A., S. Luo, and f.t.A.s. Disease Neuroimaging Initiative, *Convolutional neural networks for Alzheimer's disease detection on MRI images.* Journal of Medical Imaging, 2021. **8**(2): p. 024503-024503.
14. Hossain, T., et al. *Brain tumor detection using convolutional neural network*. in *2019 1st international conference on advances in science, engineering and robotics technology (ICASERT)*. 2019. IEEE.
15. Kushwaha, P.K., et al. *Brain Tumour Detection Using MRI Images and CNN Architecture*. in *2024 International Conference on Communication, Computer Sciences and Engineering (IC3SE)*. 2024. IEEE.
16. Alhsnony, F.H. and L. Sellami. *Advancing Breast Cancer Detection with Convolutional Neural Networks: A Comparative Analysis of MIAS and DDSM Datasets*. in *2024 IEEE 7th International Conference on Advanced Technologies, Signal and Image Processing (ATSIP)*. 2024. IEEE.
17. Hadush, S., et al., *Breast cancer detection using convolutional neural networks.* arXiv preprint arXiv:2003.07911, 2020.
18. Hamid, M.A., H.M. Mondher, and B. Ayoub. *Deep Learning CNNs for Breast Cancer Classification and Detection" Enhancing Diagnostic Accuracy in Medical Practice*. in *2024 2nd International Conference on Electrical Engineering and Automatic Control (ICEEAC)*. 2024. IEEE.
19. Jaroonroj Wongnil, A.K., Rajalida Lipikorn, *Breast cancer characterization using region-based convolutional neural network with screening and diagnostic mammogram.* JOURNAL OF ASSOCIATED MEDICAL SCIENCES (Online), 2024. **57**(3): p. 8-17.
20. Mohith K P, R.H., Anzar Iqbal, Vijay Kumar Gottipati, Gella Nagendram, V Mano Priya, Sushma Shetty, *Enhancing Breast Cancer Detection: Leveraging Convolutional Neural Networks.* International Journal for Research in Applied Science and Engineering Technology (IJRASET), 2024. **12**(7): p. 943-948.
21. Rahman, M.M., et al., *Breast cancer detection and localizing the mass area using deep learning.* Big Data and Cognitive Computing, 2024. **8**(7): p. 80.
22. Agarwal, H., et al., *Predictive Data Analysis: Leveraging RNN and LSTM Techniques for Time Series Dataset.* Procedia Computer Science, 2024. **235**: p. 979-989.
23. Ahlawat, S., *Recurrent Neural Networks*, in *Reinforcement Learning for Finance: Solve Problems in Finance with CNN and RNN Using the TensorFlow Library*. 2023, Apress: Berkeley, CA. p. 177-232.
24. Al Olaimat, M., S. Bozdag, and A.s.D.N. Initiative, *TA-RNN: an attention-based time-aware recurrent neural network architecture for electronic health records.* Bioinformatics, 2024. **40**(Supplement_1): p. i169-i179.
25. Dixit, K.K., et al. *Sequential Data Analysis in Healthcare: Predicting Disease Progression with Long Short-Term Memory Networks*. in *2023 International Conference on Artificial Intelligence for Innovations in Healthcare Industries (ICAIIHI)*. 2023. IEEE.
26. Kabila, R., et al. *Hybrid LSTM-RNN and Lion Optimization Algorithm for IoT-based Proactive Healthcare Data Management*. in *2024 International Conference on Knowledge Engineering and Communication Systems (ICKECS)*. 2024. IEEE.
27. Yadav, D., et al. *Using Long Short-Term Memory Units for Time Series Forecasting*. in *2023 2nd International Conference on Futuristic Technologies (INCOFT)*. 2023. IEEE.
28. Chen, J., et al., *LSTM-Based Prediction Model for Tuberculosis Among HIV-Infected Patients Using Structured Electronic Medical Records: A Retrospective Machine Learning Study.* Journal of Multidisciplinary Healthcare, 2024: p. 3557-3573.
29. Golchha, R., P. Khobragade, and A. Talekar. *Design of an Efficient Model for Health Status Prediction Using LSTM, Transformer, and Bayesian Neural Networks*. in *2024 International Conference on Innovations and Challenges in Emerging Technologies (ICICET)*. 2024. IEEE.
30. Jahangiri, S., et al., *A machine learning model to predict heart failure readmission: toward optimal feature set.* Frontiers in Artificial Intelligence, 2024. **7**: p. 1363226.
31. Rizinde, T., I. Ngaruye, and N. Cahill, *Machine Learning Algorithms for Predicting Hospital Readmission and Mortality Rates in Patients with Heart Failure.* African Journal of Applied Research, 2024. **10**(1): p. 316-338.
32. Tian, J., et al., *Enhancing Disease Prediction with a Hybrid CNN-LSTM Framework in EHRs.* Journal of Theory and Practice of Engineering Science, 2024. **4**(02): p. 8-14.

33. Zarghani, A., *Comparative Analysis of LSTM Neural Networks and Traditional Machine Learning Models for Predicting Diabetes Patient Readmission.* arXiv preprint arXiv:2406.19980, 2024.
34. Abbasian Ardakani, A., et al., *Interpretation of Artificial Intelligence Models in Healthcare: A Pictorial Guide for Clinicians.* Journal of Ultrasound in Medicine, 2024.
35. Ashreetha, B., et al. *Accurate Neoplasm Diagnosis with Comprehensive Machine Learning and Deep Learning Approaches*. in *2024 5th International Conference for Emerging Technology (INCET)*. 2024. IEEE.
36. Banapuram, C., A.C. Naik, and M.K. Vanteru, *A Comprehensive Survey of Machine Learning in Healthcare: Predicting Heart and Liver Disease, Tuberculosis Detection in Chest X-Ray Images.* SSRG International Journal of Electronics and Communication Engineering, 2024. **11**(5): p. 155-169.
37. Guido, R., et al., *An Overview on the Advancements of Support Vector Machine Models in Healthcare Applications: A Review.* Information, 2024. **15**(4): p. 235.
38. Viignesh, M.R., et al. *The Role of AI Techniques in Diagnosing Health Conditions with Integration of AI*. in *2024 4th International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*. 2024. IEEE.
39. Kalra, N., P. Verma, and S. Verma, *Advancements in AI based healthcare techniques with FOCUS ON diagnostic techniques.* Computers in Biology and Medicine, 2024. **179**: p. 108917.
40. Krishnamoorthy, P., et al. *Revolutionizing Medical Diagnostics: Exploring Creativity in AI for Biomedical Image Analysis*. in *2024 International Conference on Signal Processing, Computation, Electronics, Power and Telecommunication (IConSCEPT)*. 2024. IEEE.
41. M. A., M.K., Manoj Kumar M, M. A., Ms. Mahalakshmi, *A Review on Artificial Intelligence in Medicine.* International Journal of Advanced Research in Science, Communication and Technology, 2024. **7**(1): p. 268-274.
42. Sheliemina, N., *The Use of Artificial Intelligence in Medical Diagnostics: Opportunities, Prospects and Risks.* Health Economics and Management Review, 2024. **5**(2): p. 104-124.
43. Gopi, B., et al., *Distributed Technologies Using AI/ML Techniques for Healthcare Applications*, in *Social Innovations in Education, Environment, and Healthcare*. 2024, IGI Global. p. 375-396.
44. Kaur, G., H. Singh, and S. Mehta. *AI Doctors in Healthcare: A Comparative Journey through Diagnosis, Treatment, Care, Drug Development, and Health Analysis*. in *2024 Sixth International Conference on Computational Intelligence and Communication Technologies (CCICT)*. 2024. IEEE.
45. Shang, Z., et al., *Artificial Intelligence, the Digital Surgeon: Unravelling Its Emerging Footprint in Healthcare–The Narrative Review.* Journal of Multidisciplinary Healthcare, 2024: p. 4011-4022.
46. Alami, A., J. Boumhidi, and L. Chakir. *Explainability in CNN based Deep Learning models for medical image classification*. in *2024 International Conference on Intelligent Systems and Computer Vision (ISCV)*. 2024. IEEE.
47. Chattu, P., et al. *Exploring Convolution Neural Networks for Image Classification in Medical Imaging*. in *2024 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE)*. 2024. IEEE.
48. Jyotiram Sawale, P.V.B., J. M. Gandhi, Satish N. Gujar, Suresh Limkar, Samir N. Ajani, *Deep learning for image-based diagnosis : Applications in medical imaging for drug development.* Journal of Statistics and Management Systems, 2024. **27**(2): p. 201-212.
49. Dai, L., M. Zhou, and H. Liu, *Recent Applications of Convolutional Neural Networks in Medical Data Analysis*, in *Federated Learning and AI for Healthcare 5.0*. 2024, IGI Global. p. 119-131.
50. MJ, C.M.B., *Convolutional Neural Networks for Medical Image Segmentation and Classification: A Review.* Journal of Information Systems and Telecommunication (JIST), 2023. **4**(44): p. 347.
51. Eswaran, U., A. Khang, and V. Eswaran, *Applying Machine Learning for Medical Image Processing*, in *AI and IoT-Based Technologies for Precision Medicine*. 2023, IGI Global. p. 137-154.
52. Chinta, S.V., et al., *Ai-driven healthcare: A survey on ensuring fairness and mitigating bias.* arXiv preprint arXiv:2407.19655, 2024.
53. Huang, G., et al., *From explainable to interpretable deep learning for natural language processing in healthcare: How far from reality?* Computational and Structural Biotechnology Journal, 2024.
54. Yekaterina, K., *Challenges and Opportunities for AI in Healthcare.* International Journal of Law and Policy, 2024. **2**(7): p. 11-15.
55. Nasarian, E., et al., *Designing interpretable ML system to enhance trust in healthcare: A systematic review to proposed responsible clinician-AI-collaboration framework.* Information Fusion, 2024: p. 102412.
56. Solaiman, B., *The European Union's Artificial Intelligence Act and Trust: Towards an AI Bill of Rights in Healthcare?* Available at SSRN 4886518, 2024.
57. Wassenaar, P.N., et al., *The role of trust in the use of artificial intelligence for chemical risk assessment.* Regulatory Toxicology and Pharmacology, 2024. **148**: p. 105589.
58. Cheong, B.C., *Transparency and accountability in AI systems: safeguarding wellbeing in the age of algorithmic decision-making.* Frontiers in Human Dynamics, 2024. **6**: p. 1421273.

59. Bataineh, A.Q., et al. *Ethical & Legal Concerns of Artificial Intelligence in the Healthcare Sector*. in *2024 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems (ICETSIS)*. 2024. IEEE.
60. Mennella, C., et al., *Ethical and regulatory challenges of AI technologies in healthcare: A narrative review.* Heliyon, 2024.
61. Solaiman, B., *From 'AI to Law'in Healthcare: The Proliferation of Global Guidelines in a Void of Legal Uncertainty.* Barry Solaiman,'From 'AI to Law'in Healthcare: The Proliferation of Global Guidelines in a Void of Legal Uncertainty'(2023), 2023. **42**(2): p. 391-406.
62. Aquino, Y.S.J., *Making decisions: Bias in artificial intelligence and data-driven diagnostic tools.* Australian journal of general practice, 2023. **52**(7): p. 439-442.
63. Bhatia, S., A. Kumar, and S. Tandon, *Uncovering the Challenges From Algorithmic Bias Affecting the Marginalized Patient Groups in Healthcare.* Anuj and Tandon, Stuti, Uncovering the Challenges From Algorithmic Bias Affecting the Marginalized Patient Groups in Healthcare (May 30, 2024), 2024.
64. Cevik, J., et al., *Assessment of the bias of artificial intelligence generated images and large language models on their depiction of a surgeon.* ANZ Journal of Surgery, 2024. **94**(3): p. 287-294.
65. Gaonkar, N., *Exploring Bias Assessment and Strong Calibration in AI-based Medical Risk Prediction Models.* 2024.
66. Perets, O., et al., *Inherent Bias in Electronic Health Records: A Scoping Review of Sources of Bias.* medRxiv, 2024.
67. Shuford, J., *Examining Ethical Aspects of AI: Addressing Bias and Equity in the Discipline.* Journal of Artificial Intelligence General science (JAIGS) ISSN: 3006-4023, 2024. **3**(1): p. 262-280.
68. Stanley, E.A., et al., *Towards objective and systematic evaluation of bias in medical imaging AI.* arXiv preprint arXiv:2311.02115, 2023.
69. Chandio, S.A., et al., *Enhancing Trust in Healthcare: The Role of AI Explainability and Professional Familiarity.* The Asian Bulletin of Big Data Management, 2024. **4**(1): p. Science 4 (1)-21.
70. Selvaraju, R.R., et al. *Grad-cam: Visual explanations from deep networks via gradient-based localization*. in *Proceedings of the IEEE international conference on computer vision*. 2017.
71. Bassan, S., G. Amir, and G. Katz, *Local vs. Global Interpretability: A Computational Complexity Perspective.* arXiv preprint arXiv:2406.02981, 2024.
72. Kumar, A.K., et al., *Feature Contribution to an In-Depth Understanding of the Machine Learning Model Interpretation.* Przeglad Elektrotechniczny, 2024. **2024**(2).
73. Kumawat, K., *Explainable AI: Interpretable Models for Transparent Decision.* International journal of food and nutritional science, 2024.
74. Priya.V, P.B.R., *Explainable AI (XAI): Interpretable Model Architectures*. Vol. 3. 2024: IIP Series.
75. Lundberg, S.M., et al., *From local explanations to global understanding with explainable AI for trees.* Nature machine intelligence, 2020. **2**(1): p. 56-67.
76. Selvaraju, R.R., et al., *Grad-CAM: visual explanations from deep networks via gradient-based localization.* International journal of computer vision, 2020. **128**: p. 336-359.
77. Grover, V. and M. Dogra, *Challenges and Limitations of Explainable AI in Healthcare*, in *Analyzing Explainable AI in Healthcare and the Pharmaceutical Industry*. 2024, IGI Global. p. 72-85.
78. Madi, I.A.E., et al., *Exploring Explainable AI Techniques for Text Classification in Healthcare: A Scoping Review.* Digital Health and Informatics Innovations for Sustainable Health Care Systems, 2024: p. 846-850.
79. Thakur, R., *Explainable AI: Developing Interpretable Deep Learning Models for Medical Diagnosis.* International Journal For Multidisciplinary Research, 2024. **6**(4).
80. Amreen Ayesha, N.N.A., *Explainable artificial intelligence (EAI)*, in *Explainable Artificial Intelligence for Biomedical and Healthcare Applications*. 2024. p. 162-196.
81. Aziz, N.A., et al., *Explainable AI in Healthcare: Systematic Review of Clinical Decision Support Systems.* medRxiv, 2024: p. 2024.08. 10.24311735.
82. Saurabh Singhal, A.K.S., Akhilesh Kumar Singh, Abhinav Pandey, Avinash Kumar Sharma, *Industry-Specific Applications of AI and ML*, in *Advances in systems analysis, software engineering, and high performance computing book series*. 2024, IGI Global. p. 110-124.
83. Mamalakis, M., et al., *The Explanation Necessity for Healthcare AI.* arXiv preprint arXiv:2406.00216, 2024.
84. Michail, M., Héloïse, de, Vareilles., Graham, K., Murray., Píetro, Lió., John, Suckling, *The Explanation Necessity for Healthcare AI.* Cornell University, 2024.
85. Adla, P., Vasavi, Chithanuru., Posham, Uppamma., R., Vishnukumar, *Exploring Explainable AI in Healthcare: Challenges and Future Directions*, in *Analyzing Explainable AI in Healthcare and the Pharmaceutical Industry*. 2024, IGI Global. p. 35.
86. Korgialas, C., E. Pantraki, and C. Kotropoulos. *Interpretable Face Aging: Enhancing Conditional Adversarial Autoencoders with Lime Explanations*. in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2024. IEEE.
87. Rashid, M., et al. *Using Stratified Sampling to Improve LIME Image Explanations*. in *Proceedings of the AAAI Conference on Artificial Intelligence*. 2024.

88. Sharma, N., et al., *A novel dataset and local interpretable model-agnostic explanations (LIME) for monkeypox prediction.* Intelligent Decision Technologies, 2023(Preprint): p. 1-12.
89. Mane, D., et al., *Unlocking Machine Learning Model Decisions: A Comparative Analysis of LIME and SHAP for Enhanced Interpretability.* Journal of Electrical Systems, 2024. **20**(2s): p. 1252-1267.
90. Choi, H.-W. and S. Abdirayimov, *Demonstrating the Power of SHAP Values in AI-Driven Classification of Marvel Characters.* Journal of Multimedia Information System, 2024. **11**(2): p. 167-172.
91. Goldwasser, J. and G. Hooker, *Provably Stable Feature Rankings with SHAP and LIME.* arXiv preprint arXiv:2401.15800, 2024.
92. Hu, L. and K. Wang, *Computing SHAP Efficiently Using Model Structure Information.* arXiv preprint arXiv:2309.02417, 2023.
93. Huang, X. and J. Marques-Silva, *Updates on the Complexity of SHAP Scores.*
94. Kelodjou, G., et al. *Shaping Up SHAP: Enhancing Stability through Layer-Wise Neighbor Selection*. in *Proceedings of the AAAI Conference on Artificial Intelligence*. 2024.
95. Mohammed, B., A. Anouar, and B. Nadjia, *Grad-CAM Guided preprocessing and convolutional neural network for efficient mammogram images classification.* Informatica, 2024. **47**(10).
96. Joshi, D., et al. *Revealing Advanced Brain Tumour Detection: An In-Depth Study Leveraging Grad CAM Interpretability*. in *International Joint Conference on Advances in Computational Intelligence*. 2022. Springer.
97. Muntasir, F., A. Datta, and S. Mahmud. *Interpreting Multiclass Lung Cancer from CT Scans using Grad-CAM on Lightweight CNN Layers*. in *2024 6th International Conference on Electrical Engineering and Information & Communication Technology (ICEEICT)*. 2024. IEEE.
98. Rao, B.N., S.K. Sabut, and R. Mishra. *Deep Learning Approach to Predict Intracerebral Hemorrhage and Grad-Cam Visualization on CT Images*. in *2024 International Conference on Recent Advances in Electrical, Electronics, Ubiquitous Communication, and Computational Intelligence (RAEEUCCI)*. 2024. IEEE.
99. Shrikumar, A., P. Greenside, and A. Kundaje. *Learning important features through propagating activation differences*. in *International conference on machine learning*. 2017. PMlR.
100. Abhyankar, G. and R. Raman. *Decision Tree Analysis for Point-of-Care Ultrasound Imaging: Precision in Constrained Healthcare Settings*. in *2024 International Conference on Inventive Computation Technologies (ICICT)*. 2024. IEEE.
101. Jammal, M., et al., *Impact on clinical guideline adherence of Orient-COVID, a CDSS based on dynamic medical decision trees for COVID19 management: a randomized simulation trial.* arXiv preprint arXiv:2407.11205, 2024.
102. Modak, M., et al. *Improving Diagnostic Accuracy: A Deep Dive into Random Forest Optimization for Clinical Data*. in *2024 3rd International Conference on Artificial Intelligence For Internet of Things (AIIoT)*. 2024. IEEE.
103. Morales Rodríguez, D., M. Pegalajar Cuellar, and D.P. Morales, *On the fusion of soft-decision-trees and concept-based models.* Available at SSRN 4402768, 2023.
104. Zhu, X., et al., *A Development of Fuzzy Rule-based Regression Models through Using Decision Trees.* IEEE Transactions on Fuzzy Systems, 2024.
105. Stubbin, A., et al., *The Limits of Perception: Analyzing Inconsistencies in Saliency Maps in XAI.* arXiv preprint arXiv:2403.15684, 2024.
106. Mahmud, T., et al., *An explainable AI paradigm for Alzheimer's diagnosis using deep transfer learning.* Diagnostics, 2024. **14**(3): p. 345.
107. Mansouri, D., et al. *Explainable AI Framework for Alzheimer's Diagnosis Using Convolutional Neural Networks*. in *2024 IEEE 7th International Conference on Advanced Technologies, Signal and Image Processing (ATSIP)*. 2024. IEEE.
108. Reena, L., Jyoti, Wadmare., Sunita, Patil., Ganesh, Wadmare., Darshan, Patil, *Transparent precision: Explainable AI empowered breast cancer recommendations for personalized treatment.* IAES International Journal of Artificial Intelligence, 2024. **13**(3): p. 2694-2694.
109. Reena R Lokare, J.W., Sunita Patil, Ganesh Wadmare, Darshan Patil, *Transparent precision: Explainable AI empowered breast cancer recommendations for personalized treatment.* IAES International Journal of Artificial Intelligence, 2024. **13**(3): p. 2694-2694.
110. Kedar, M.M., *Exploring the Effectiveness of SHAP over other Explainable AI Methods.* Indian Scientific Journal Of Research In Engineering And Management, 2024.
111. Poonia, R.C. and H.A. Al-Alshaikh, *Ensemble approach of transfer learning and vision transformer leveraging explainable AI for disease diagnosis: An advancement towards smart healthcare 5.0.* Computers in Biology and Medicine, 2024. **179**: p. 108874.
112. Ribeiro, M.T., S. Singh, and C. Guestrin. *" 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*. 2016.
113. Loh, H.W., et al., *Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022).* Computer Methods and Programs in Biomedicine, 2022. **226**: p. 107161.
114. Lundberg, S., *A unified approach to interpreting model predictions.* arXiv preprint arXiv:1705.07874, 2017.
115. Doshi-Velez, F. and B. Kim, *Towards a rigorous science of interpretable machine learning.* arXiv preprint arXiv:1702.08608, 2017.

116. Samek, W., *Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models.* arXiv preprint arXiv:1708.08296, 2017.
117. Keller, J., et al., *Exercise leads to metabolic changes associated with improved strength and fatigue in people with MS.* Annals of Clinical and Translational Neurology, 2021. **8**(6): p. 1308-1317.
118. Mehrabi, N., et al., *A survey on bias and fairness in machine learning.* ACM computing surveys (CSUR), 2021. **54**(6): p. 1-35.
119. Gianfrancesco, M.A., et al., *Potential biases in machine learning algorithms using electronic health record data.* JAMA internal medicine, 2018. **178**(11): p. 1544-1547.
120. Rajkomar, A., J. Dean, and I. Kohane, *Machine learning in medicine.* New England Journal of Medicine, 2019. **380**(14): p. 1347-1358.
121. Watson, D.S., et al., *Clinical applications of machine learning algorithms: beyond the black box.* Bmj, 2019. **364**.
122. He, J., et al., *The practical implementation of artificial intelligence technologies in medicine.* Nature medicine, 2019. **25**(1): p. 30-36.
123. Al-Rubaie, M. and J.M. Chang, *Privacy-preserving machine learning: Threats and solutions.* IEEE Security & Privacy, 2019. **17**(2): p. 49-58.
124. Christoph, M., *Interpretable machine learning: A guide for making black box models explainable*. 2020: Leanpub.
125. Parikh, R.B., S. Teeple, and A.S. Navathe, *Addressing bias in artificial intelligence in health care.* Jama, 2019. **322**(24): p. 2377-2378.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.