Explainable Artificial Intelligence and Interpretable Machine Learning in Healthcare: A Survey

Abstract

Explainable Artificial Intelligence (XAI) and Interpretable Machine Learning (IML) are critical in transforming healthcare by addressing the opacity of traditional AI models. This survey highlights the importance of XAI in enhancing transparency, trust, and efficacy in medical applications, emphasizing its role in improving decision-making and patient outcomes. XAI techniques, such as SHAP and LIME, provide insights into AI-driven decisions, fostering trust among healthcare professionals by elucidating complex model behaviors. The survey identifies challenges, including model complexity, data privacy, and lack of standardization, while advocating for robust frameworks and domain-specific techniques to enhance interpretability. User-centric approaches and interdisciplinary collaboration are emphasized as essential for developing effective XAI systems that align with stakeholder needs. Ethical and regulatory considerations are also addressed, underscoring the need for comprehensive frameworks to ensure responsible AI deployment in healthcare. The survey concludes that the integration of XAI into healthcare systems is pivotal for realizing the full potential of AI technologies in improving patient care and outcomes. Ongoing advancements in XAI promise to revolutionize healthcare, offering pathways to more transparent, trustworthy, and effective AI-driven solutions. Future research should focus on developing adaptive explanation systems, enhancing real-time inference capabilities, and exploring interdisciplinary collaborations to advance XAI in healthcare.

1 Introduction

1.1 Importance of XAI and IML in Healthcare

Explainable Artificial Intelligence (XAI) and Interpretable Machine Learning (IML) are pivotal in transforming healthcare technologies by mitigating the opacity associated with traditional AI models, often criticized for their 'black box' nature. These methodologies enhance transparency and interpretability, enabling healthcare professionals to comprehend the rationale behind AI-driven decisions, which is critical in high-stakes medical contexts where erroneous predictions can have dire consequences. By fostering trust and facilitating informed decision-making, XAI and IML are essential for the effective integration of AI into clinical practices, ensuring clinicians can confidently rely on AI outputs. Ongoing research into various XAI techniques, such as Local Interpretable Model-Agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP), paves the way for safer and more effective AI applications in healthcare, ultimately improving patient outcomes [1, 2, 3, 4, 5]. This transformation is crucial in healthcare, where transparency is vital due to the high stakes involved in patient safety and outcomes.

In clinical environments, the adoption of XAI techniques significantly enhances decision-making and patient care by fostering trust and understanding. XAI techniques are specifically designed to address the transparency challenges posed by the black-box nature of deep neural networks, which is essential in critical applications such as healthcare, autonomous vehicles, and military operations,

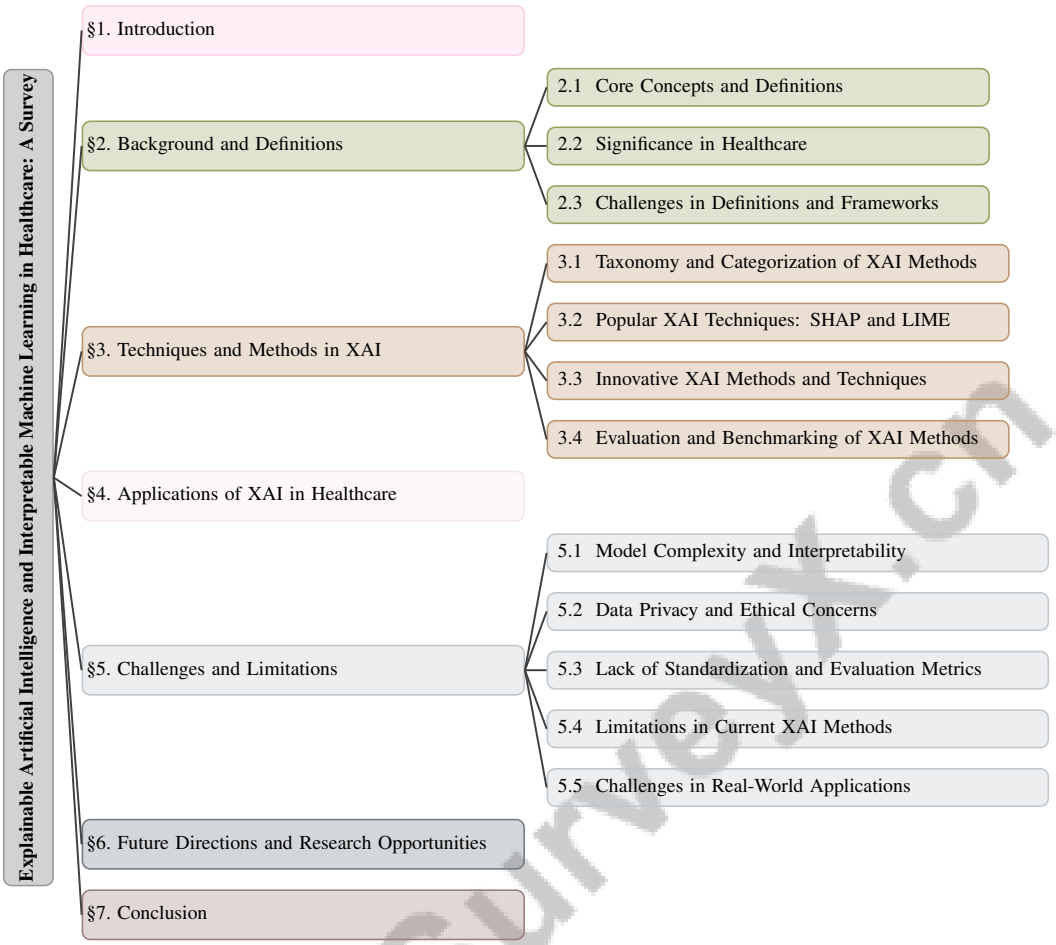


Figure 1: chapter structure

where decisions can profoundly impact human lives. By providing interpretable and comprehensible explanations of AI decision-making processes, XAI enhances trust and addresses ethical and judicial concerns, thereby bolstering the reliability and accountability of AI systems in these vital sectors [6, 7]. The focus on transparency and accountability in safety-critical systems underscores the importance of XAI in healthcare.

Moreover, XAI offers a pragmatic and naturalistic understanding, ensuring that explanations are accessible to stakeholders, thus enhancing the reliability and trustworthiness of AI models. The ability of XAI to clarify the significance of features in machine learning models further solidifies its role in augmenting decision-making processes in healthcare. As advancements in AI technologies progress, their role in enhancing the interpretability and trustworthiness of AI models in healthcare becomes increasingly vital. AI systems, often perceived as "black boxes," can create uncertainties in medical decision-making. XAI methods are essential for elucidating AI decision-making processes, promoting transparency and confidence among healthcare professionals. By employing various interpretability techniques, XAI not only addresses concerns of bias and opacity but also encourages broader adoption of AI solutions in critical healthcare applications, ultimately improving patient outcomes [8, 1, 9, 3, 10].

XAI and IML also play a critical role in addressing misconceptions and enhancing the design and effectiveness of AI explanations through human-centered approaches, crucial for various stakeholders. In high-stakes applications, where trust is paramount, XAI ensures transparency, particularly in safety-critical systems like clinical diagnostics, by providing insights into how AI models arrive at predictions and addressing domain experts' concerns regarding model reliability. This is achieved through various XAI techniques that enhance interpretability and foster trust among healthcare professionals, ultimately facilitating the responsible integration of AI in healthcare settings [11, 12].

AI-generated, for reference only.

XAI and IML are integral to bridging the gap between complex AI models and their practical applications in healthcare, thereby improving patient outcomes and fostering trust in AI-driven healthcare solutions. As healthcare technologies evolve, the integration of XAI and IML will be crucial for ensuring that AI systems in medical settings are not only effective in predictive capabilities but also transparent and trustworthy, addressing critical concerns regarding the 'black box' nature of AI decision-making that could impact patient outcomes. This integration aims to enhance clinicians' understanding of AI-generated recommendations, fostering greater trust and facilitating the safe and responsible deployment of these technologies in healthcare [3, 1, 2, 11].

1.2 Motivation for the Survey

The motivation for this survey stems from the pressing need to enhance the interpretability and trustworthiness of AI models in healthcare, where incorrect predictions can severely impact patient care [1]. The survey aims to address the urgent requirement for explainable AI (XAI) systems, focusing on how AI applications can become more transparent and trustworthy for users [13]. A significant challenge lies in ensuring that AI models, particularly those utilized in digital twins for Remaining Useful Life (RUL) prediction, are explainable, interpretable, and trustworthy. This involves overcoming the limitations of existing models in providing understandable predictions and ensuring transparency in decision-making processes [14].

A primary challenge identified is the disconnect between technical XAI approaches and users' actual needs and goals, often resulting in explanations that fail to enhance understanding or trust [15]. This disconnect highlights the importance of aligning XAI methods with user expectations to improve the effectiveness of AI explanations. Furthermore, the survey addresses the lack of trust in machine learning models due to misconceptions surrounding XAI, which impede their deployment in high-risk domains such as healthcare [16].

By synthesizing existing research on the utility of XAI in human-AI decision-making, the survey seeks to clarify ambiguous findings regarding the impact of XAI on user performance. This initiative is essential for underscoring the significance of trustworthy AI systems in healthcare, particularly in high-stakes environments where transparency and interpretability are critical. As the healthcare sector increasingly integrates AI technologies, concerns about the opacity of these systems and potential biases in their predictions necessitate a focus on Explainable Artificial Intelligence (XAI). By employing various interpretability techniques, healthcare professionals can better understand AI-driven decisions, fostering trust and facilitating the responsible deployment of AI in medical contexts where accurate diagnosis and treatment are paramount [8, 1, 11, 9, 3].

1.3 Structure of the Survey

This survey is systematically organized to provide a comprehensive overview of Explainable Artificial Intelligence (XAI) and Interpretable Machine Learning (IML) within the context of healthcare, particularly focusing on traditional medicine applications. The paper is structured to guide readers through interconnected sections that build upon each other, offering a holistic understanding of the subject matter.

The survey begins with an **Introduction**, highlighting the significance of XAI and IML in healthcare, followed by a detailed discussion on the **Importance of XAI and IML in Healthcare**. This section underscores the transformative potential of these technologies in enhancing transparency and trust in AI systems. The **Motivation for the Survey** elaborates on the driving factors behind this research, emphasizing the critical need for trustworthy AI in healthcare applications.

Next, the paper delves into the **Background and Definitions**, defining core concepts such as Explainable AI, Interpretable Machine Learning, SHAP, and Model Interpretability within the healthcare and traditional medicine context. The **Techniques and Methods in XAI** section explores various XAI methodologies, focusing on their application in healthcare and providing a taxonomy and categorization of existing methods, including popular techniques like SHAP and LIME.

The **Applications of XAI in Healthcare** section examines specific use cases and the impact of XAI on transparency and trust in medical applications, discussing its role in safety-critical domains and the benefits of improved user interaction. The **Challenges and Limitations** section identifies

obstacles to implementing XAI in healthcare, such as model complexity, data privacy, and the lack of standardization.

In the **Future Directions and Research Opportunities**, the survey outlines potential research trajectories, advocating for the development of robust XAI frameworks and domain-specific explainability techniques. It emphasizes the significance of interdisciplinary collaboration and the necessity of addressing ethical and regulatory considerations in the development and implementation of XAI. This is crucial as diverse stakeholders—including those in healthcare, finance, and technology—have varying expectations and requirements for transparency and accountability in AI systems. By fostering interdisciplinary dialogue, researchers can better align XAI approaches with stakeholders' needs, ensuring that ethical guidelines and regulatory standards are effectively integrated to prevent misuse and enhance trust in AI technologies [17, 18, 8].

Finally, the **Conclusion** synthesizes key findings and reinforces the importance of XAI in healthcare. This structured approach ensures that the survey provides a clear roadmap for readers, facilitating an understanding of the multifaceted nature of XAI and its implications across various domains, as emphasized by Longo et al. in their organization of XAI research into fields like healthcare, finance, and education [19].The following sections are organized as shown in Figure 1.

2 Background and Definitions

2.1 Core Concepts and Definitions

Explainable Artificial Intelligence (XAI) and Interpretable Machine Learning (IML) are essential for developing transparent and accountable AI systems, especially in critical sectors like healthcare [20]. XAI methodologies aim to demystify machine learning models, addressing their inherent opacity by providing tailored explanations for diverse stakeholders, including developers and decision-makers, which is crucial for clinical validation and fostering trust in AI-driven decisions [2]. In contrast, IML emphasizes intrinsic model transparency, reducing the need for supplementary explanatory tools, which is vital in healthcare settings where understanding model decisions is paramount [2]. The distinction between XAI's post-hoc explanations and IML's inherently understandable models is significant.

Key XAI techniques include SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations). SHAP applies cooperative game theory to assess feature contributions to predictions, offering comprehensive insights into model behavior [21]. LIME approximates models locally with interpretable ones to clarify specific predictions [2]. Despite their advantages, challenges such as 'inflated explanations' in non-additive predictive models persist [2].

Model interpretability is crucial for the effective and trusted deployment of AI systems, particularly in high-stakes environments like healthcare [22]. Crafting meaningful explanations for diverse audiences remains challenging, emphasizing the need for tailored explainability solutions in complex domains like healthcare [23]. Porcedda et al.'s taxonomy highlights key concepts like transparency, interpretability, and understandability [24]. The distance-restricted explanations (DRE) method offers a novel approach, ensuring explanations remain valid within specified proximities using adversarial robustness tools [25]. These efforts underscore the ongoing pursuit of formalizing criteria for explainability in AI, addressing the current lack of clear formalization in IML.

2.2 Significance in Healthcare

The integration of Explainable Artificial Intelligence (XAI) and Interpretable Machine Learning (IML) into healthcare is crucial due to the complexity of AI models, often considered 'black boxes.' Such opacity poses significant challenges for healthcare professionals who rely on these systems for critical decision-making, as unclear AI predictions can adversely affect patient outcomes. XAI techniques enhance the interpretability of AI outputs, fostering trust among clinicians and ensuring the underlying reasoning of AI systems is accessible and comprehensible. By addressing AI complexities in high-stakes medical environments, XAI contributes to safer and more effective healthcare practices, ultimately improving patient care quality [8, 1, 2, 3, 4]. The opaque nature of these models necessitates developing robust, efficient, and fair models that enhance trust and usability while addressing data bias and privacy concerns.

XAI techniques provide transparent explanations, fostering trust among healthcare practitioners and patients, and facilitating the seamless integration of AI systems into clinical workflows [23]. The comprehensibility of XAI-generated explanations ensures they are meaningful and actionable for stakeholders, bridging the gap between complex AI models and user needs [15]. This is particularly significant in high-stakes applications where decisions directly impact patient outcomes [1].

The importance of XAI is further highlighted by its ability to clarify variable importance in machine learning models, which is critical for reliable decision-making in healthcare [21]. Inaccurate assessments of variable importance can lead to misleading conclusions, emphasizing the necessity for interpretable models that healthcare professionals can easily understand [22]. This is especially relevant in traditional medicine, where incorporating prior knowledge into interpretability methods enhances the relevance and applicability of AI explanations across diverse healthcare settings [26].

Moreover, the complexity of logical reasoning presents a significant challenge in achieving explainability, particularly when managing numerous inputs in machine learning models [25]. This complexity underscores the need for tailored explainability solutions, particularly in healthcare, where verifying model predictions is challenging yet essential for robust performance and transparent decision-making [27]. As AI technologies advance, the roles of XAI and IML in ensuring transparency, trust, and ethical integrity will become increasingly vital, facilitating the successful deployment of AI systems in both conventional and traditional medical practices [26].

Definitions of transparency, interpretability, and explainability are crucial for understanding AI systems, highlighting their importance in the healthcare context [24]. These definitions provide a framework for evaluating AI systems, ensuring they meet necessary standards for effective and trustworthy deployment in healthcare environments, where the stakes are inherently high.

2.3 Challenges in Definitions and Frameworks

Establishing clear definitions and frameworks for Explainable Artificial Intelligence (XAI) and Interpretable Machine Learning (IML) faces significant challenges due to the absence of universally accepted definitions and evaluation criteria. This lack of consensus leads to confusion and ineffective communication between AI systems and users, as there are no common definitions, metrics, or benchmarks for what constitutes an explanation in AI systems [16]. The complexity of deep learning models exacerbates this issue, as their intricate architectures hinder interpretability and complicate the identification of biases, particularly those arising from demographic disparities in training data [28].

Current XAI methods often inadequately address multicollinearity, resulting in misleading interpretations and a lack of robustness in explanations [29]. This problem is intensified by the widespread acceptance of non-formal XAI methods lacking rigor, which can mislead users and undermine trust in AI systems [16]. Additionally, the absence of standardized methods for evaluating model performance in dynamic environments complicates establishing effective benchmarks for XAI [30].

The challenges in defining and evaluating XAI are compounded by the lack of agreed-upon evaluation criteria, protocols, and frameworks for assessing the interpretability and explainability of AI models. Despite various proposals, no gold standard for evaluation methods exists, complicating the assessment of explainability approaches in XAI. This is further complicated by the common practice of not evaluating the effectiveness of XAI methods, leading to potential misinterpretations of their reliability in healthcare applications. Moreover, biases introduced by imputation methods can distort feature importance measures, resulting in misleading interpretations [16].

In light of these challenges, there is an urgent need for standardized metrics and frameworks that address the diverse needs of stakeholders in healthcare and beyond. Ongoing efforts to establish clear definitions and frameworks for XAI and IML are crucial for advancing the field, as they aim to tackle the black-box nature of AI systems. By providing structured explanations that encompass essential dimensions such as format, completeness, accuracy, and currency, these frameworks enhance the reliability and meaningfulness of AI outputs. This, in turn, fosters trust and usability in real-world applications, particularly in critical sectors like healthcare, finance, and autonomous driving, where transparency and accountability are paramount. Furthermore, a comprehensive understanding of XAI techniques and their implications on user behavior will guide future research and development, ultimately leading to more effective AI systems that align with user needs and societal expectations [31, 32, 12, 33].

3 Techniques and Methods in XAI

Category	Feature	Method
Popular XAI Techniques: SHAP and LIME	Feature Importance Assessment	MIP[34], ShapleyVIC[21]
Innovative XAI Methods and Techniques	User-Centric Customization	GM-CW[35]
	Performance Optimization Techniques	DRE[25]
	Feature and Data Interactions	XAI-MDS[36], XAI-CI[37]
	Pattern-Based Interpretability	DLIPM[38]

Table 1: This table provides a comprehensive overview of various Explainable Artificial Intelligence (XAI) techniques and methods, categorizing them into popular techniques such as SHAP and LIME, and innovative methods focusing on user-centric customization, performance optimization, feature interactions, and pattern-based interpretability. It highlights the diversity of approaches used to enhance AI interpretability and transparency, particularly in the context of healthcare applications.

In Explainable Artificial Intelligence (XAI), methodologies are pivotal for enhancing the interpretability and transparency of AI systems. This section explores various techniques and methods that augment understanding and trust in AI applications. The following subsection categorizes XAI methods, providing a structured overview of diverse approaches for achieving explainability in artificial intelligence. Table 1 categorizes and summarizes key XAI techniques and methods, illustrating their roles in enhancing the interpretability and transparency of AI systems. Additionally, Table 3 presents a comparative overview of key XAI techniques, elucidating their distinct characteristics and contributions to model interpretability and transparency. Figure 2 illustrates the hierarchical structure of these techniques and methods, categorizing them into taxonomy and methods, including popular techniques like SHAP and LIME, as well as innovative methods and evaluation frameworks. This figure highlights key dimensions, method classifications, and applications, offering a comprehensive overview of XAI's landscape in enhancing AI interpretability and transparency, particularly in healthcare contexts.

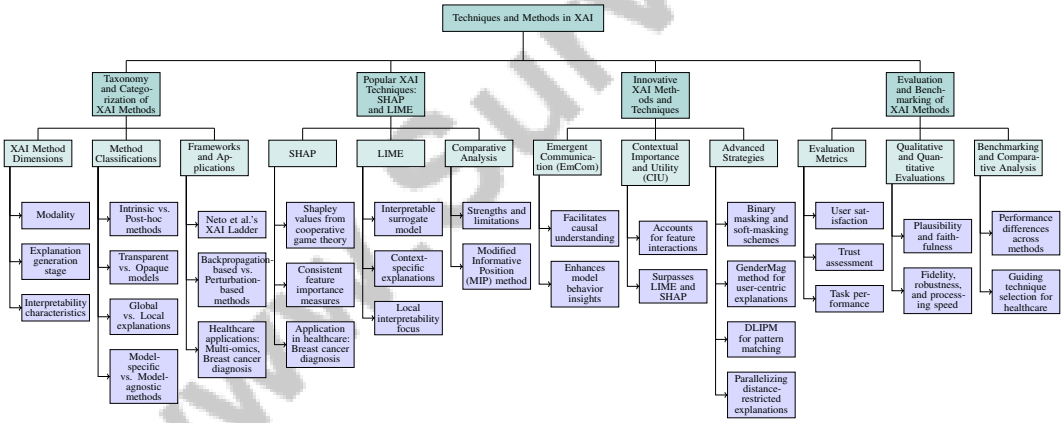


Figure 2: This figure illustrates the hierarchical structure of techniques and methods in Explainable Artificial Intelligence (XAI), categorizing them into taxonomy and methods, popular techniques like SHAP and LIME, innovative methods, and evaluation frameworks. It highlights key dimensions, method classifications, and applications, providing a comprehensive overview of XAI's landscape in enhancing AI interpretability and transparency, particularly in healthcare contexts.

3.1 Taxonomy and Categorization of XAI Methods

The taxonomy of XAI methods is crucial for understanding strategies that improve model interpretability and transparency. As illustrated in Figure 3, this figure categorizes XAI methods into types based on inherent interpretability, levels of explanation, and application contexts. It highlights the primary distinction between intrinsic and post-hoc methods: intrinsic methods are inherently interpretable, offering transparency without extra tools, while post-hoc methods generate explanations post-training, often utilizing feature importance and perturbation-based techniques.

Further classification distinguishes transparent from opaque models, emphasizing models that provide clear decision-making insights [12]. Global and local explanations form another key classification;

AI-generated, for reference only.

global explanations offer insights into overall model behavior, whereas local explanations focus on individual predictions [39]. Model-specific methods leverage unique properties of particular models, while model-agnostic methods apply across various models, ensuring flexibility and broad applicability [40].

Neto et al.'s XAI Ladder framework categorizes XAI methods by interpretability and causal understanding, underscoring the importance of causal insights in enhancing XAI rigor [28]. Additionally, categorization into backpropagation-based and perturbation-based methods further enriches the understanding of their application in various contexts, including healthcare.

In medical applications, distinguishing between model-agnostic and model-specific frameworks enhances AI decision understanding, as seen in multi-omics and breast cancer diagnosis [40]. The integration of user-centric methodologies reflects XAI's evolving landscape, emphasizing alignment with user expectations.

The taxonomy of XAI methods provides a structured overview of the existing landscape, highlighting the interconnectedness of various approaches and their relevance to different domains. This comprehensive categorization is essential for advancing XAI, facilitating AI systems' development that offers reliable and meaningful explanations crucial for trust, transparency, and usability across diverse applications, especially in safety-critical environments [33, 41, 24, 42, 32].

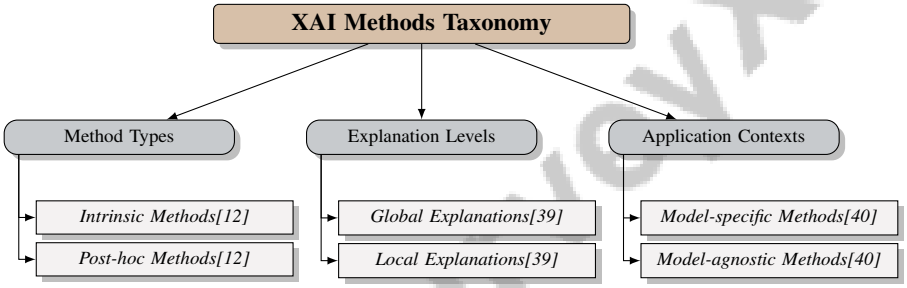


Figure 3: This figure illustrates the taxonomy of Explainable AI (XAI) methods, categorizing them into types based on inherent interpretability, levels of explanation, and application contexts. It highlights intrinsic and post-hoc methods, global and local explanations, and the distinction between model-specific and model-agnostic methods.

3.2 Popular XAI Techniques: SHAP and LIME

SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) are key XAI techniques enhancing model interpretability across domains, notably in healthcare. Both are post-hoc approaches elucidating feature importance post-training [43].

SHAP employs Shapley values from cooperative game theory to quantify each feature's contribution to predictions, providing consistent feature importance measures across models [21]. This is particularly valuable in healthcare, such as breast cancer diagnosis, offering robust local explanations by quantifying individual feature influences [44]. However, SHAP may struggle to distinguish relevant from irrelevant features in certain contexts, as seen in comparisons with methods like Optimal Trees [22].

LIME approximates a model locally with an interpretable surrogate model, generating context-specific explanations for individual predictions [9]. It excels in local interpretability, crucial in domains where understanding specific decision instances is vital. Nonetheless, LIME may lack global consistency, limiting comprehensive model insights [44].

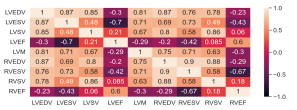
The comparative analysis of SHAP and LIME underscores their distinct strengths and limitations in elucidating model behavior. Both techniques advance XAI by offering transparent, interpretable explanations that facilitate AI integration into critical domains like healthcare [43]. The development of innovative methods, such as the Modified Informative Position (MIP) method, further enriches the landscape by providing novel approaches contrasting with traditional techniques like SHAP and LIME [34].

Evaluation method/Axiom/Metric	Method proposer and/or example authors who employed them
Sensitivity analysis	Ahlstrom et al. (2020)
Example images	Ribeiro, Singh, and Guestrin (2016)
SAT/Model counting	Ignatiev et al. (2018); Narodytska et al. (2019); Ignatiev, Narodytska, and Marques-Silva (2019)
Correlation, simplefulness, and complexity	Chu, Lee, and Shafiei (2019)
Concordance, continuity	Montavon (2018)
Fidelity	Lakkaraju, Arsov, and Bastani (2020)
Valid features	Ribeiro, Singh, and Guestrin (2016)
Post-hoc accuracy	Chen et al. (2018); Bhatt, Ravikumar, and Moura (2019); Xie and Ermon (2019); Bai et al. (2020)
Perturbation analysis for vision	Zeiler and Fergus (2014)
ROAR	Hooker et al. (2019)
Localization on Time Series	Schlegel et al. (2019)
Implementation invariance, sensitivity	Sundararajan, Taly, and Yan (2017)
Input invariance	Kindermans et al. (2019)
Simulated users	Ribeiro, Singh, and Guestrin (2016)
Amazon Mechanical Turk users	Ribeiro, Singh, and Guestrin (2016); Chen et al. (2018)
In-depth Interviews	Hoffman et al. (2018)

(a) The table presents a list of various evaluation methods and metrics used in the field of machine learning, along with the authors who have employed them.[45]

```
explainer = LimeTabularExplainer(
    convert_to_lime_format(X_train,
        categorical_names).
    values,
    mode="classification",
    feature_names=X_train.columns.tolist(),
    categorical_names=categorical_names,
    categorical_features=categorical_features=categorical_names.keys(),
    discretize_continuous=True,
    random_state=42)
```

(b) A Python code snippet for creating a LimeTabularExplainer object[46]



(c) The image is a heat map displaying the correlation coefficients between various cardiac measurements.[34]

Figure 4: Examples of Popular XAI Techniques: SHAP and LIME

As shown in Figure 4, understanding and interpreting decisions made by complex machine learning models is crucial in XAI. This example introduces SHAP and LIME, two widely recognized techniques for achieving interpretability. Figure 4 illustrates these techniques, with the first subfigure presenting a comprehensive table of evaluation methods and metrics in machine learning, highlighting diverse approaches to evaluating model interpretability. The second subfigure provides a Python code snippet demonstrating the creation of a LimeTabularExplainer object, essential for generating local explanations of model predictions using the LIME framework. The third subfigure displays a heat map illustrating correlation coefficients between various cardiac measurements, offering insights into feature relationships crucial for model interpretation. Together, these examples underscore the importance of SHAP and LIME in enhancing AI systems' transparency and accountability by providing clear insights into decision-making processes [45, 46, 34].

3.3 Innovative XAI Methods and Techniques

Innovative methods in XAI are reshaping AI applications, especially in healthcare, where transparency and interpretability are critical. Integrating emergent communication (EmCom) into AI systems, as proposed by Perrett et al., enhances XAI by facilitating causal understanding of outputs, providing deeper insights into model behavior [47]. This is crucial in healthcare, where understanding causal pathways leading to predictions can significantly impact clinical decision-making.

The CIU (Contextual Importance and Utility) method introduced by Knapi et al. offers nuanced explanations by accounting for feature interactions and context, surpassing traditional methods like LIME and SHAP in providing comprehensive insights [36]. This method's ability to consider feature interplay is particularly beneficial in complex healthcare scenarios, where multiple variables influence patient outcomes.

Apicella et al. propose innovative strategies, such as binary masking and soft-masking schemes, to merge input data with explanations, enhancing AI model interpretability [37]. These strategies integrate data-driven insights with explanatory frameworks, vital for healthcare professionals relying on AI for accurate diagnostics.

The application of the GenderMag method, explored by Hamid et al., aims to improve XAI systems by addressing users' problem-solving styles, tailoring explanations to diverse user needs [35]. This user-centric approach ensures AI explanations are not only technically sound but also accessible and relevant to healthcare practitioners with varying expertise.

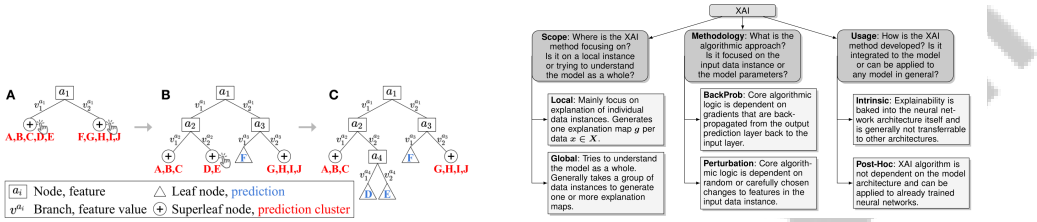
Neto et al. emphasize developing methods that enhance interpretability through causal reasoning and counterfactual explanations, highlighting future research trajectories that could revolutionize AI evaluation and trust in healthcare [28]. Such methods are essential for understanding potential outcomes of different clinical interventions, supporting informed decision-making.

The development of DLIPM (Deep Learning Interpretability through Pattern Matching) represents a significant advancement in leveraging deep learning techniques to automatically learn complex patterns from large datasets, addressing limitations of traditional linear models [38]. This innovation is particularly relevant in healthcare, where discerning intricate patterns in patient data can lead to more accurate and personalized treatment plans.

AI-generated, for reference only.

The introduction of novel algorithms for parallelizing the computation of distance-restricted explanations, as described by Izza et al., significantly enhances performance compared to existing methods [25]. This advancement is crucial for real-time healthcare applications, where rapid and reliable explanations are necessary for timely clinical decisions.

These emerging techniques underscore the dynamic evolution of XAI, highlighting innovative methods' critical role in enhancing transparency, trustworthiness, and applicability of AI systems in healthcare. As the healthcare field evolves with advanced analytical techniques and diverse data availability, advancements in XAI are set to enhance AI systems' transparency and reliability. This will improve AI integration in clinical settings and ensure medical practitioners can confidently utilize these technologies for high-quality patient care, addressing interpretability and potential bias concerns in AI-driven predictions [8, 10].



(a) Feature Clustering in a Decision Tree[48] (b) XAI (explainable artificial intelligence)[6]

Figure 5: Examples of Innovative XAI Methods and Techniques

As shown in Figure 5, innovative methods and techniques are paramount for enhancing the transparency and interpretability of complex AI models. The examples in Figure 5 showcase two distinct approaches in this domain. The first example, "Feature Clustering in a Decision Tree," visually demonstrates how decision trees can effectively cluster features, highlighting the process through an insightful diagram. This technique emphasizes splitting a decision tree into branches based on feature values, aiding in understanding the model's decision-making process. The second example, represented by a comprehensive flowchart, delves into the broader concept of XAI itself, systematically breaking down its multifaceted nature into three core sections: Scope, Methodology, and Usage. Each section is meticulously detailed, offering a granular view of how XAI can be applied and understood. Together, these examples underscore the innovative strides in XAI, illustrating both specific techniques and overarching frameworks that contribute to making AI systems more transparent and interpretable [48, 6].

3.4 Evaluation and Benchmarking of XAI Methods

Benchmark	Size	Domain	Task Format	Metric
XAI-Elec[49]	9,000,000	Electrification	Classification	Accuracy, MAE
XAI-Metrics[50]	4,000	Computer Vision	Image Classification	Faithfulness Correlation, Max-Sensitivity
GTE[45]	2,600,000	Energy Consumption	Classification	C-of-ED, Second Correct
LEAF[51]	1,000	Health Risk Assessment	Binary Classification	Local Fidelity, Reiteration Similarity
LIME[46]	142,193	Weather Prediction	Classification	Accuracy, ROC-AUC
tlos[52]	1,235	Medical Imaging	Survival Analysis	C-index, IBS
XAI-Benchmark[53]	768	Health	Binary Classification	Accuracy, F1
ML-Cost-Pred[54]	986	Medical Insurance	Cost Prediction	R-squared, RMSE

Table 2: This table presents a comprehensive overview of various benchmarks utilized for evaluating Explainable Artificial Intelligence (XAI) methods across multiple domains. Each benchmark is characterized by its size, domain of application, task format, and the specific metrics used for assessment. These benchmarks are instrumental in understanding the performance and applicability of XAI methods in diverse fields, particularly in critical sectors like healthcare.

Evaluating and benchmarking XAI methods is essential for effective implementation in healthcare, where transparency and trust are paramount. An effective evaluation framework must include diverse metrics, such as user satisfaction, trust assessment, and task performance, to comprehensively assess trade-offs between interpretability and accuracy [13]. The survey by Sadeghi et al. highlights

comparative analyses of various XAI methods, emphasizing their effectiveness and applicability in healthcare, delineating strengths and weaknesses of different approaches [1].

Evaluating XAI methods involves both qualitative and quantitative metrics. Qualitative evaluations focus on the plausibility and faithfulness of explanations, ensuring alignment with user expectations and enhancing trust. Quantitative metrics, such as fidelity, robustness, and processing speed, provide objective measures of explanation quality, facilitating assessment of how well explanations reflect the model's true behavior [13]. These metrics are crucial for determining explanation generation efficiency, ensuring timely and reliable insights in healthcare settings.

Developing robust benchmarks for evaluating XAI methods is vital for identifying suitable techniques for specific applications. Such benchmarks reveal significant performance differences across methods, underscoring the importance of model choice in interpreting feature sensitivities and handling complex scenarios, such as multicollinearity [1]. Table 2 provides a detailed overview of representative benchmarks employed in the evaluation and benchmarking of XAI methods, highlighting their relevance and application across different domains. The comparative analysis in Sadeghi et al.'s survey provides valuable insights into the varying capabilities of these methods, guiding the selection of appropriate techniques for healthcare applications.

Ongoing efforts to evaluate and benchmark XAI methods are crucial for advancing the field and ensuring AI systems provide reliable, meaningful explanations. Enhancing the integration of AI systems into critical sectors, such as healthcare, is vital for ensuring transparency and interpretability, essential for building trust among medical practitioners and improving decision-making processes. As AI technologies advance, particularly through XAI techniques, they address concerns regarding model opacity and potential biases, fostering greater confidence in AI-driven medical diagnoses. By employing various interpretability methods, healthcare professionals can better understand AI outputs, leading to more informed and reliable patient care [8, 11, 9, 3, 10].

Feature	SHAP	LIME	Emergent Communication (EmCom)
Explanation Type	Post-hoc	Post-hoc	Intrinsic
Model Compatibility	Model-agnostic	Model-agnostic	Healthcare-specific
Key Strength	Consistent Feature Importance	Local Interpretability	Causal Understanding

Table 3: This table provides a comparative analysis of three prominent Explainable Artificial Intelligence (XAI) techniques: SHAP, LIME, and Emergent Communication (EmCom). It highlights their explanation types, model compatibility, and key strengths, emphasizing their roles in enhancing interpretability and transparency, particularly in healthcare applications.

4 Applications of XAI in Healthcare

The integration of Explainable Artificial Intelligence (XAI) in healthcare is increasingly pivotal for enhancing the interpretability and trustworthiness of AI systems. As reliance on AI for clinical decision-making grows, understanding these systems' decision-making processes becomes crucial, especially when patient care and outcomes are significantly impacted. This section examines XAI's role in fostering transparency and trust in medical applications, highlighting its contribution to improving healthcare delivery.

4.1 Enhancing Transparency and Trust in Medical Applications

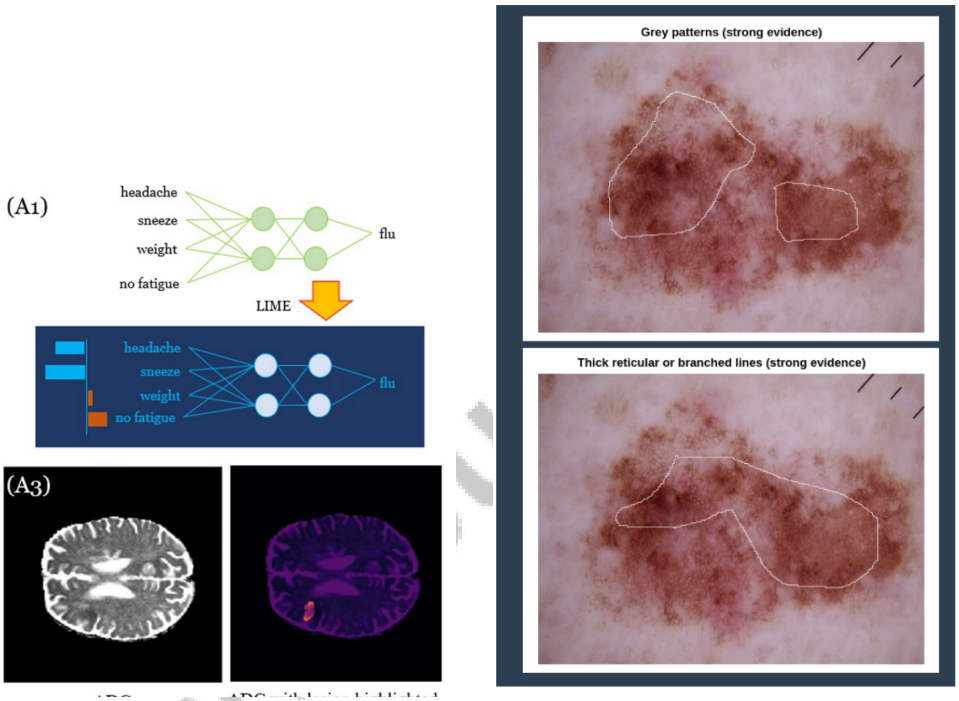
XAI plays a crucial role in enhancing transparency and trust within healthcare systems, essential for the widespread adoption of AI technologies. By providing interpretable insights, XAI instills confidence among healthcare professionals and patients [20]. Its application in multi-omics analysis aids in biomarker identification and clinical decision-making, underscoring its importance in enhancing model interpretability [40].

In breast cancer diagnostics, XAI methods like SHAP, LIME, and Grad-CAM have improved diagnostic accuracy while providing interpretable explanations. These techniques, integrated with advanced models, enhance the detection and classification of breast cancer in various imaging modalities, addressing implementation challenges in clinical settings and ultimately improving

patient outcomes [55, 56, 57, 44, 58]. These advancements highlight XAI's transformative potential by ensuring AI systems are effective, transparent, and reliable.

Empirical evidence shows that XAI techniques elucidate AI models' decision-making processes, fostering transparency and enabling critical assessments in high-stakes environments where inaccuracies can lead to severe consequences [8, 3, 1, 2]. XAI not only improves diagnostic accuracy but also enhances decision-making processes, fostering trust in AI technologies.

As AI technologies evolve, integrating XAI is crucial for maintaining transparency and trust, enhancing clinical adoption. The continuous evolution of XAI techniques is expected to significantly improve AI's integration and effectiveness in delivering high-quality patient care. By ensuring AI models are interpretable and transparent, XAI builds trust among healthcare professionals, addressing the critical need for accountability in AI applications and reinforcing the importance of understanding AI systems' reasoning, ultimately fostering a reliable and effective healthcare delivery system [1, 2, 11, 3, 5].



(a) Deep Learning and Brain Imaging: A Comparative Study[5]

(b) Skin Lesion Analysis[55]

Figure 6: Examples of Enhancing Transparency and Trust in Medical Applications

As illustrated in Figure 6, XAI's integration in healthcare significantly enhances transparency and trust. The studies highlighted demonstrate the comparative analysis of brain imaging techniques and XAI's application in dermatology for skin lesion diagnosis. The first study showcases neural networks' ability to interpret complex variables and diagnose conditions through MRI data. The second emphasizes XAI's role in accurately distinguishing skin lesions. Together, these examples underscore XAI's potential to foster a more transparent, trustworthy, and interpretable healthcare environment, enhancing decision-making and patient outcomes [5, 55].

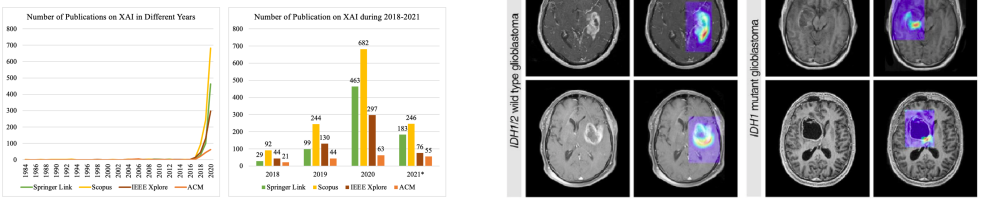
4.2 Applications in Safety-Critical Domains

Deploying XAI in safety-critical healthcare domains is essential for patient safety and reliable AI-driven decisions. In high-stakes environments, XAI's transparent insights foster trust among healthcare professionals and patients [2]. Integrating XAI into intensive care monitoring and surgical decision support enables clinicians to understand and validate AI-generated recommendations, reducing adverse outcomes [20].

XAI is crucial for identifying and mitigating potential errors in AI systems, vital for maintaining patient safety. By providing clear explanations for predictions, XAI facilitates early detection of anomalies and biases, allowing timely intervention to prevent harm [27]. This capability is critical in emergency medicine and critical care, where rapid and accurate decision-making is essential [2].

Moreover, XAI enhances AI systems' accountability by providing transparent explanations, allowing professionals to trace AI-driven decisions' reasoning [26]. This traceability ensures compliance with ethical standards and regulatory requirements, paramount in healthcare settings where patient safety is prioritized [2].

XAI's context-specific explanations adapt AI systems to meet safety-critical domains' unique needs. By tailoring explanations to clinical scenarios, XAI enhances AI systems' relevance and applicability, improving their effectiveness in supporting patient care [2]. As AI technologies advance, integrating XAI into safety-critical domains will ensure AI systems are effective, safe, and trustworthy, enhancing patient safety and outcomes.



(a) Number of Publications on XAI in Different Years[42]

(b) Comparison of MRI and PET scans in glioblastoma patients with wild-type and mutant IDH1/2[59]

Figure 7: Examples of Applications in Safety-Critical Domains

As shown in Figure 7, XAI's application in safety-critical domains, particularly healthcare, is gaining significant attention. The increasing number of XAI publications from 1984 to 2021 underscores its potential to enhance decision-making in safety-critical environments. A notable application is the analysis of medical imaging, such as MRI and PET scans, for glioblastoma patients. By comparing scans of patients with wild-type and mutant IDH1/2, XAI provides deeper insights into disease characteristics and progression, crucial for tailored treatment plans and improved patient outcomes, illustrating XAI's transformative impact on managing complex medical conditions [42, 59].

4.3 Benefits of XAI in Improving User Interaction

XAI significantly enhances user interaction and decision-making in healthcare by offering clear explanations of AI predictions. This transparency is vital in high-stakes medical environments, where incorrect predictions can have severe consequences. By utilizing various XAI methodologies—features-oriented methods, global and local explainability techniques, and human-centric approaches—clinicians can better understand AI systems' reasoning, fostering trust and informed decisions. This interpretability improves AI applications' safety and efficacy, addressing healthcare professionals' diverse needs and ensuring AI tools are reliable and user-friendly [1, 33, 2, 11, 3]. By elucidating AI-driven decisions' reasoning, XAI fosters deeper understanding and trust among healthcare professionals, enhancing clinical decision-making.

A key advantage of XAI is its ability to improve model performance through strategic explanations that guide feature selection and model tuning [37]. This approach enhances AI model accuracy and optimizes relevance to specific healthcare applications, improving user experience. By aligning AI models with healthcare professionals' needs and expectations, XAI facilitates more intuitive interactions, enabling informed decisions based on reliable and comprehensible data.

Furthermore, XAI techniques enhance user interaction by providing context-specific explanations tailored to various healthcare scenarios' unique requirements. This adaptability ensures AI systems are technically proficient and meet users' needs with varying expertise levels, providing personalized explanations and on-demand supplementary information that enhance trust, transparency, understandability, usability, and fairness [60, 33]. By bridging the gap between complex AI models and practical

AI-generated, for reference only.

applications, XAI plays a crucial role in improving user satisfaction and engagement in healthcare settings.

As AI technologies advance, integrating XAI into healthcare systems will enhance AI-driven solutions' effectiveness and user-friendliness. XAI aims to make algorithms transparent and understandable, addressing trust issues with "black box" models, crucial in healthcare, where decisions impact patient outcomes. By improving interpretability and fostering trust among healthcare professionals, XAI can facilitate more reliable clinical decision-making and ultimately support better patient care [3, 11, 1, 2]. Ongoing XAI advancements promise to enhance the interaction between healthcare professionals and AI systems, leading to improved decision-making and patient outcomes.

5 Challenges and Limitations

The exploration of challenges and limitations in Explainable Artificial Intelligence (XAI) within healthcare necessitates an understanding of the relationship between model complexity and interpretability. This interplay is pivotal as it affects the efficacy and trustworthiness of AI systems in clinical settings. The following subsection delves into these complexities, emphasizing the need for a balance between sophisticated model design and clear, interpretable outcomes comprehensible to healthcare professionals.

5.1 Model Complexity and Interpretability

The trade-offs between model complexity and interpretability are significant in healthcare, where transparent AI systems are essential. Complex models, like deep neural networks, offer high predictive performance but are often criticized for their 'black box' nature, complicating trust and understanding in clinical environments [2]. This opacity is problematic in healthcare, where interpretability is crucial for trust and safe AI deployment [2].

Challenges include accurately characterizing the Rashomon set of nearly optimal models, especially in high-dimensional or correlated data contexts, affecting both complexity and interpretability [21]. Balancing accuracy with interpretability is difficult, mainly due to the lack of reliable operational data for training [14]. This complexity necessitates a unified framework to effectively categorize and compare XAI methods.

Experiments by Dunn et al. show interpretable methods like Optimal Trees outperform black-box methods like XGBoost in feature selection, highlighting the importance of choosing models that balance complexity and interpretability [22]. However, many studies fall short in addressing diverse user needs and evaluating XAI techniques in real-world applications [15].

The computational complexity of some XAI methods poses challenges for real-time clinical applications, where timely decision-making is critical [25]. Addressing these challenges requires nuanced approaches considering performance and interpretability trade-offs, ensuring AI systems are effective, transparent, and trustworthy [2].

5.2 Data Privacy and Ethical Concerns

Integrating XAI into healthcare systems raises data privacy and ethical challenges that must be addressed for responsible deployment. Regulatory frameworks impose strict requirements on AI systems, potentially limiting the adoption of complex models [61]. While necessary for safety and accountability, these regulations can hinder innovation in healthcare AI [19].

Key obstacles include transparency, trust, and data privacy, which are critical ethical considerations [62]. Users express concerns about data handling and AI system transparency [17], highlighting the need for XAI methods that communicate effectively with end-users [63]. Variability in XAI explanations and potential for incorrect interpretations due to incomplete data or feature dependencies further complicate these challenges [64].

Data privacy issues are heightened when handling sensitive medical data, such as administrative datasets lacking comprehensive clinical data, impacting predictive accuracy [63]. The complexity of implementing adaptive XAI systems that respond to diverse user contexts raises questions about scalability and practical applicability in real-world healthcare [62].

The static nature of data, which may not reflect ongoing changes in patient conditions, poses another limitation for XAI in healthcare [19]. This can lead to outdated insights and reduce AI system effectiveness in dynamic environments. Human annotation reliance, costly and time-consuming, limits current XAI studies' scalability [39].

Despite advancements, comprehensive evaluation metrics and frameworks to measure XAI effectiveness are lacking [16]. This gap underscores the need for robust validation frameworks to ensure XAI's reliability and trustworthiness in clinical settings. Explainable AI methods are essential for accountability and trust in high-stakes medical applications [17].

Addressing data privacy and ethical concerns requires establishing standardized guidelines prioritizing patient safety, trust, and ethical integrity. Ongoing efforts to address these issues in XAI for healthcare are crucial for responsible AI integration into clinical practice, ensuring AI systems are effective, transparent, and trustworthy [19].

5.3 Lack of Standardization and Evaluation Metrics

The absence of standardized evaluation metrics in XAI presents significant challenges in assessing explanation quality and effectiveness, particularly in healthcare settings with diverse user needs. Current studies often lack comprehensive evaluation mechanisms, limiting XAI methods' applicability in real-world scenarios [65]. This lack of standardization leads to inconsistencies in explanation assessment, resulting in ambiguous interpretations of their utility [44].

Existing benchmarks frequently lack thorough evaluation metric analysis, resulting in unclear reliability and effectiveness. Variability in tasks and experimental setups complicates XAI method comparisons, leading to inconsistent findings [30]. The lack of consensus on good evaluation methods for explainability approaches complicates establishing effective benchmarks.

Efforts to address these challenges include developing benchmarks using ground-truth explanations (GTE) to evaluate XAI method accuracy, promoting systematic advancement. Focusing solely on individual XAI implementations without comprehensive comparisons can result in biased insights. This limitation underscores the need for holistic evaluation approaches incorporating a systematic taxonomy of explainability dimensions—such as transparency, interpretability, completeness, complexity, and understandability—to facilitate a clearer understanding of XAI's effectiveness and address challenges associated with AI systems' black-box nature [32, 24].

Ongoing efforts to establish standardized metrics and evaluation frameworks are crucial for advancing XAI and ensuring AI systems provide reliable explanations. These efforts are vital for effective AI technology integration into healthcare, where transparency and interpretability are paramount. By employing XAI techniques, healthcare professionals can better understand AI-driven predictions, fostering clinician trust and enhancing decision-making processes. This transparency is essential to mitigate concerns about potential biases and inaccuracies in AI models, ensuring medical practitioners can confidently rely on AI systems in critical situations, ultimately leading to improved patient outcomes [8, 1, 11, 3, 10].

5.4 Limitations in Current XAI Methods

Current XAI methods face limitations, particularly in traditional medicine, where clear and trustworthy explanations are crucial. A significant challenge is reliance on post-hoc explanations, which, while widely used, can be misleading and fail to capture full model behavior complexity, especially in high-stakes environments like healthcare [43]. This issue is exacerbated by the subjective nature of explanations, which can vary significantly depending on the user's perspective, complicating trust in AI systems [23].

Many XAI methods remain black boxes, providing explanations lacking depth and comprehensibility needed by clinicians [1]. The complexity of biological systems and deep learning models further complicates understanding AI-driven insights in healthcare, leading to oversimplified interpretations that do not adequately reflect medical decision-making intricacies [40].

Current studies often fall short in addressing non-expert user needs, resulting in explanations that are technically sound but not practically useful [23]. This gap highlights the importance of developing XAI methods catering to diverse user needs and contexts, ensuring explanations are accurate, accessi-

ble, and actionable [13]. Moreover, the rapid evolution of XAI methods creates uncertainty among researchers regarding appropriate techniques for specific questions, complicating XAI integration into traditional medicine [66].

Reliance on human evaluation in XAI studies may introduce biases, and existing taxonomies may not cover all explainability aspects, indicating limitations in current methods [24]. Dataset quality dependency can affect XAI result generalizability, posing challenges for diverse healthcare applications [30].

Addressing these limitations requires developing sophisticated XAI methods providing reliable, context-specific explanations tailored to traditional medicine's unique needs. This includes integrating social and cultural factors into the explanation process to enhance healthcare application effectiveness [26]. As XAI evolves, overcoming these challenges is essential for ensuring AI systems deliver meaningful and trustworthy explanations that enhance traditional medicine decision-making.

5.5 Challenges in Real-World Applications

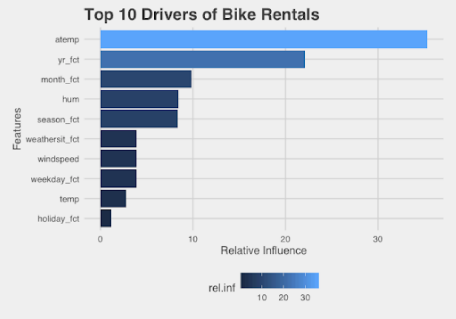
Applying XAI techniques in real-world healthcare settings presents several challenges for effective deployment. One major challenge is the need for extensive, well-annotated datasets, essential for training and validating XAI models for high accuracy [4]. The scarcity of such datasets in healthcare can hinder XAI system development and implementation, limiting their ability to provide reliable and meaningful explanations.

Integrating diverse AI methodologies into a cohesive framework poses another challenge in real-world applications. Healthcare data heterogeneity and the need for interoperability between AI systems complicate seamless XAI technique integration into clinical workflows [4]. This complexity can lead to difficulties in ensuring AI-driven insights are actionable and relevant to healthcare practitioners, impacting XAI's overall effectiveness in enhancing clinical decision-making.

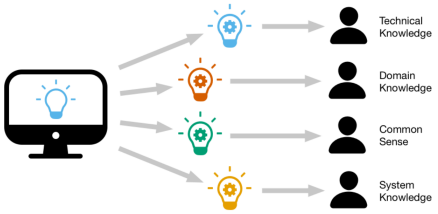
The dynamic nature of healthcare environments, characterized by rapidly evolving medical knowledge and practices, necessitates continuous adaptation and updating of XAI models to remain relevant and accurate. The necessity for continuous refinement and validation of XAI models poses a significant logistical challenge, requiring substantial resources and specialized expertise to ensure XAI systems' reliability in practical applications. This ongoing process is crucial not only for enhancing model performance and generalization but also for aligning AI systems with user needs and regulatory standards, highlighting the complex interplay between explainability and system performance. Furthermore, the effectiveness of XAI methods in improving model properties can vary widely based on factors such as the specific model, dataset, and explanation techniques employed, necessitating careful consideration and strategic planning in their implementation [67, 68].

Successful XAI implementation in healthcare also depends on addressing diverse stakeholder needs, including clinicians, patients, and regulatory bodies. Ensuring XAI explanations are accessible and comprehensible to non-expert users is crucial for fostering trust and facilitating AI technology adoption in clinical practice. To effectively address Explainable Artificial Intelligence (XAI) complexities, developing user-centric methods that enhance clarity and usability while ensuring compliance with ethical and regulatory standards is essential. This involves prioritizing key dimensions such as explanation format, completeness, accuracy, and currency, significantly influencing user trust, transparency, understandability, usability, and fairness. Furthermore, XAI method evaluation should focus on their impact on human decision-making, considering context-dependent explanations and potential explainability-system performance trade-offs. By integrating these considerations, a comprehensive framework aligning XAI with user needs and societal expectations can be created, ultimately advancing AI-supported systems' effectiveness [67, 33, 69].

As shown in Figure 8, in real-world applications, integrating data-driven insights is met with challenges and limitations. The "Top 10 Drivers of Bike Rentals" example presents a bar chart highlighting features influencing bike rental patterns and their relative importance, underscoring the challenge of identifying and quantifying various factors' impact in predictive modeling, where understanding each feature's weight is crucial for accurate decision-making. The "Knowledge Flow Diagram" depicts knowledge dissemination from a central source to diverse stakeholders, emphasizing the difficulty of effectively communicating technical, domain, and other knowledge types to varied audiences with distinct expertise and needs. Together, these examples encapsulate real-world application chal-



(a) Top 10 Drivers of Bike Rentals[59]



(b) Knowledge Flow Diagram[70]

Figure 8: Examples of Challenges in Real-World Applications

lenges, from accurately modeling and interpreting data to ensuring effective knowledge transfer and comprehension across sectors [59, 70].

6 Future Directions and Research Opportunities

The advancement of Explainable Artificial Intelligence (XAI) hinges on establishing foundational elements that ensure AI systems are effective and trustworthy. This section outlines key components crucial for XAI's progression, emphasizing the need for robust frameworks that enhance transparency and interpretability, particularly in healthcare where clarity and accountability are critical.

6.1 Development of Robust XAI Frameworks

Robust XAI frameworks are pivotal for ensuring AI systems in healthcare are transparent and clinically viable. Future research should focus on adaptive explanation systems tailored to individual user needs, fostering engagement and trust [14]. This involves integrating user requirements into XAI design, ensuring explanations are accessible and relevant across diverse healthcare stakeholders. Research should also advance algorithms for managing multi-fidelity data, real-time inference, and uncertainty quantification for noisy datasets [14]. Expanding formal methods to encompass a wider range of machine learning models and improving scalability is crucial for addressing misconceptions and enhancing interpretability. Refining evaluation metrics to tackle emerging challenges will facilitate the creation of standardized benchmarks, strengthening XAI methods across sectors. Future efforts should also enhance heuristics for scaling distance-restricted explanation methods and explore advanced hardware to improve performance, thereby increasing XAI's real-world applicability [25]. Investigating user interaction with explanations and the effects of limiting explanation numbers will advocate for frameworks balancing comprehensibility and information overload [24]. Addressing these priorities will significantly enhance AI technology integration in healthcare, benefiting clinicians and patients by ensuring AI systems are effective, transparent, and trustworthy.

6.2 Domain-Specific Explainability Techniques

Developing domain-specific explainability techniques is essential for improving AI system interpretability in healthcare, where medical practice complexities necessitate tailored explanation approaches. Intelligent Decision Assistance systems, as proposed by Schemmer et al., offer a promising method for incorporating human oversight into AI processes, enhancing explainability and aligning AI insights with clinical expertise [71]. These systems encourage collaboration between human professionals and AI technologies, facilitating decision-making that leverages both human intuition and computational precision. The specificity of medical data and clinical workflow intricacies require explainability techniques finely tuned to various medical domains, such as radiology, oncology, or cardiology. By addressing healthcare's unique challenges, these XAI techniques provide actionable insights that enhance AI system effectiveness and trustworthiness in clinical practice. Clear model behavior explanations foster greater confidence among healthcare professionals, facilitating responsible AI integration into diagnostics and treatment processes [72, 8]. Incorporating domain-specific

explainability techniques is crucial for navigating healthcare's ethical and regulatory challenges, as these methods enhance transparency and trust in AI systems. By ensuring AI models produce interpretable outputs, healthcare professionals can better comprehend automated decision-making, vital for compliance with industry standards and best practices. This approach promotes accountability and fairness, mitigating risks associated with opaque decision-making in high-stakes medical contexts [8, 1, 72, 11, 3]. Tailored mechanisms for accountability and transparency addressing healthcare professionals, patients, and regulatory bodies' specific needs are critical for fostering trust and AI technology acceptance in healthcare. As AI integration in healthcare progresses, developing domain-specific explainability techniques becomes increasingly vital for enhancing AI system effectiveness and ensuring transparent and comprehensible decision-making processes for healthcare professionals. Given the high stakes of medical decision-making, where erroneous predictions can have dire consequences, establishing trust in AI applications through XAI methods is imperative. This trust is essential for clinicians to confidently rely on AI-driven insights in patient care, ultimately leading to broader adoption and improved healthcare outcomes [8, 1, 2, 11, 3]. This approach promises to enhance AI-driven decision interpretability, ultimately improving patient outcomes and advancing medical practice.

6.3 Integration of User-Centric Approaches

Integrating user-centric approaches in XAI is crucial for enhancing AI systems' usability and accessibility, especially in healthcare, where diverse user needs must be met. User-centric methodologies prioritize designing explanations tailored to end-user requirements and preferences, ensuring AI-driven insights are comprehensible and actionable [73]. By focusing on user perspectives, these methods facilitate the development of XAI systems that are intuitive and effective in supporting clinical decision-making. Incorporating user-centric methodologies requires engaging stakeholders throughout the development process to understand their expectations and challenges. This collaboration is essential for designing explanations that align with healthcare professionals' cognitive and practical needs, thereby enhancing trust and acceptance of AI technologies [54]. By closely collaborating with end-users, developers can create meaningful and contextually relevant explanations that improve the overall user experience. Furthermore, integrating user-centric approaches can mitigate barriers to data access hindering effective XAI system deployment. Collaborating with stakeholders, such as insurance providers, can facilitate better data sharing and integration, leading to more robust and reliable AI models [54]. This collaboration is crucial for equipping XAI systems with high-quality data, enabling them to deliver accurate and trustworthy explanations meeting healthcare practitioners' needs. The integration of user-centric approaches in XAI is essential for developing AI systems demonstrating technical proficiency while resonating with users' values, expectations, and informational needs. Research indicates effective XAI should encompass dimensions such as trust, transparency, understandability, usability, and fairness, providing explanations that are complete, accurate, and relevant. By prioritizing user experiences and involving users in the development process, AI systems can be designed to offer tailored, non-technical explanations addressing users' specific queries, including data usage concerns and recommendation rationale. Shifting from a purely technical focus to a more human-centered design perspective will enhance user engagement and foster a deeper understanding of AI functionalities [60, 32, 15, 33]. As AI technologies evolve, prioritizing user-centric design will be crucial for enhancing XAI systems' usability and effectiveness, ultimately improving patient outcomes and advancing AI integration into healthcare practice.

6.4 Interdisciplinary Collaboration and Emerging Trends

Interdisciplinary collaboration is vital for advancing XAI research, particularly in healthcare, where diverse expertise can address complex AI system challenges [74]. This collaborative approach merges insights from human-computer interaction, cognitive science, and technical AI development, ensuring XAI systems are robust and user-friendly [17]. Aligning XAI methods with end-user needs enhances AI technologies' practical applicability in real-world settings [68]. The significance of collaboration among researchers, clinicians, and AI experts is underscored by the need to advance XAI in healthcare, where the stakes are high, and AI technologies' impact is profound [3]. Future research should emphasize empirical investigations of stakeholder needs, context-sensitive explainability approaches, and interdisciplinary collaborations to enhance understanding and satisfaction of desiderata [17]. This includes understanding user contexts and creating explanations facilitating actionable understanding, critical for improving XAI system design and effectiveness [15]. Emerging trends in XAI

research, such as developing domain-specific techniques and exploring new explanation frameworks, underscore the field's dynamic nature [68]. These trends highlight the necessity for innovative approaches enhancing Interpretable Machine Learning (IML) methods' practical applicability across various domains. By fostering collaboration across disciplines and focusing on innovative technique development, researchers can ensure AI systems are effective, transparent, and trustworthy, ultimately enhancing their adoption and impact across sectors.

6.5 Ethical and Regulatory Considerations

Ethical and regulatory considerations in XAI research are paramount, particularly as AI systems become more integrated into healthcare applications. Addressing these considerations involves navigating the performance-explainability trade-off, essential for ensuring AI systems are effective and transparent [75]. Future research should focus on developing nuanced regulatory frameworks considering XAI policies' long-term effects on market dynamics and consumer trust, thereby enhancing AI technologies' reliability and acceptance [76]. Exploring unified frameworks for explanations in machine learning is vital for addressing bias and fairness issues, especially in regulated industries where AI decisions can have profound implications [77]. Such frameworks must consider XAI's social and technical aspects, promoting a holistic approach aligning with societal values and ethical standards [32]. In healthcare, the ethical application of feature importance methods, such as MIP, raises significant concerns regarding interpretation accuracy and its impact on medical decisions [34]. The AI Act, GDPR, and AI Liability Directive complicate this landscape by imposing stringent requirements on explainability and accountability, necessitating robust compliance strategies to uphold patient privacy and data security [61]. Developing personalized explanation interfaces for decision support systems is another area of interest, as these interfaces can enhance user understanding and trust by tailoring explanations to individual needs [20]. This aligns with the need for XAI methods adaptable to various domains and stakeholder requirements, ensuring effective communication of explanations while respecting patients' privacy rights. Addressing ethical and regulatory considerations in XAI research is essential for fostering trust and accountability in AI systems. This requires ongoing collaboration among researchers, policymakers, and stakeholders to develop comprehensive frameworks guiding AI technologies' ethical deployment in healthcare and beyond. Future research should continue exploring these dimensions, focusing on validating evaluation methods and identifying gaps in current practices [67].

7 Conclusion

The integration of Explainable Artificial Intelligence (XAI) into healthcare systems is essential for enhancing transparency, trust, and efficacy in AI-driven medical applications. This survey highlights XAI's transformative potential in mitigating the opacity of traditional AI models, thereby enabling more informed decision-making among healthcare professionals [55]. By clarifying AI-driven decisions, XAI improves the interpretability of complex models and fosters trust and collaboration between humans and AI systems [78]. The literature discusses frameworks of nudges and boosts that present promising strategies for designing XAI systems that enhance human-AI collaboration.

Moreover, XAI's critical role in improving patient outcomes is evident through the development of domain-specific techniques and innovative frameworks that enhance interpretability and inform effective medical interventions [79]. Emphasizing user-centric approaches ensures that AI explanations are accessible and meaningful, thereby bolstering trust and collaborative performance in healthcare settings [80].

The necessity of integrating XAI into e-health systems is particularly significant for creating user-centric and accessible digital health solutions for older adults [81]. This integration is vital for ensuring that AI technologies are effective, inclusive, and adaptable to the diverse needs of various user groups.

The proposed theoretical frameworks offer a coherent structure for discussing and designing explanations in XAI, especially in biomedicine, where the importance of faithfulness and plausibility is paramount [82]. These frameworks are crucial in guiding future research and development efforts, ensuring that XAI systems are robust, reliable, and aligned with ethical and regulatory standards in healthcare practice.

References

[1] Zahra Sadeghi, Roohallah Alizadehsani, Mehmet Akif Cifci, Samina Kausar, Rizwan Rehman, Priyakshi Mahanta, Pranjal Kumar Bora, Ammar Almasri, Rami S. Alkhawaldeh, Sadiq Hussain, Bilal Alatas, Afshin Shoeibi, Hossein Moosaei, Milan Hladik, Saeid Nahavandi, and Panos M. Pardalos. A brief review of explainable artificial intelligence in healthcare, 2023.

[2] Tim Hulsen. Explainable artificial intelligence (xai): concepts and challenges in healthcare. *AI*, 4(3):652–666, 2023.

[3] Subrato Bharati, M. Rubaiyat Hossain Mondal, and Prajoy Podder. A review on explainable artificial intelligence for healthcare: Why, how, and when?, 2023.

[4] Al Amin, Kamrul Hasan, Saleh Zein-Sabatto, Deo Chimba, Imtiaz Ahmed, and Tariqul Islam. An explainable ai framework for artificial intelligence of medical things, 2024.

[5] Erico Tjoa and Cuntai Guan. A survey on explainable artificial intelligence (xai): Towards medical xai, 2020.

[6] Arun Das and Paul Rad. Opportunities and challenges in explainable artificial intelligence (xai): A survey, 2020.

[7] Md. Ariful Islam, M. F. Mridha, Md Abrar Jahin, and Nilanjan Dey. A unified framework for evaluating the effectiveness and enhancing the transparency of explainable ai methods in real-world applications, 2024.

[8] Devam Dave, Het Naik, Smiti Singhal, and Pankesh Patel. Explainable ai meets healthcare: A study on heart disease dataset, 2020.

[9] Daniel Sierra-Botero, Ana Molina-Taborda, Mario S. Valdés-Tresanco, Alejandro Hernández-Arango, Leonardo Espinosa-Leal, Alexander Karpenko, and Olga Lopez-Acevedo. Selecting interpretability techniques for healthcare machine learning models, 2024.

[10] Tin Lai. Interpretable medical imagery diagnosis with self-attentive transformers: A review of explainable ai for health care, 2023.

[11] Akshat Dubey, Zewen Yang, and Georges Hattab. Ai readiness in healthcare through storytelling xai, 2024.

[12] Prashant Gohel, Priyanka Singh, and Manoranjan Mohanty. Explainable ai: current status and future directions, 2021.

[13] David Gunning. Explainable artificial intelligence (xai). *Defense advanced research projects agency (DARPA), nd Web*, 2(2):1, 2017.

[14] Kazuma Kobayashi and Syed Bahauddin Alam. Explainable, interpretable trustworthy ai for intelligent digital twin: Case study on remaining useful life, 2024.

[15] Q Vera Liao and Kush R Varshney. Human-centered explainable ai (xai): From algorithms to user experiences. *arXiv preprint arXiv:2110.10790*, 2021.

[16] Joao Marques-Silva. Disproving xai myths with formal methods – initial results, 2023.

[17] Markus Langer, Daniel Oster, Timo Speith, Holger Hermanns, Lena Kästner, Eva Schmidt, Andreas Sesing, and Kevin Baum. What do we want from explainable artificial intelligence (xai)? – a stakeholder perspective on xai and a conceptual model guiding interdisciplinary xai research, 2021.

[18] Patrick Hall, Navdeep Gill, and Nicholas Schmidt. Proposed guidelines for the responsible use of explainable machine learning, 2019.

[19] Luca Longo, Mario Brcic, Federico Cabitza, Jaesik Choi, Roberto Confalonieri, Javier Del Ser, Riccardo Guidotti, Yoichi Hayashi, Francisco Herrera, Andreas Holzinger, Richard Jiang, Hassan Khosravi, Freddy Lecue, Gianclaudio Malgieri, Andrés Páez, Wojciech Samek, Johannes Schneider, Timo Speith, and Simone Stumpf. Explainable artificial intelligence (xai) 2.0: A manifesto of open challenges and interdisciplinary research directions, 2023.

[20] Christian Meske and Enrico Bunde. Transparency and trust in human-ai-interaction: The role of model-agnostic explanations in computer vision-based decision support, 2020.

[21] Yilin Ning, Marcus Eng Hock Ong, Bibhas Chakraborty, Benjamin Alan Goldstein, Daniel Shu Wei Ting, Roger Vaughan, and Nan Liu. Shapley variable importance clouds for interpretable machine learning, 2021.

[22] Jack Dunn, Luca Mingardi, and Ying Daisy Zhuo. Comparing interpretability and explainability for feature selection, 2021.

[23] Szymon Bobek, Paloma Korycińska, Monika Krakowska, Maciej Mozolewski, Dorota Rak, Magdalena Zych, Magdalena Wójcik, and Grzegorz J. Nalepa. User-centric evaluation of explainability of ai with and for humans: a comprehensive empirical study, 2024.

[24] Riccardo Porcedda. Interpretability is not explainability: New quantitative xai approach with a focus on recommender systems in education, 2023.

[25] Yacine Izza, Xuanxiang Huang, Antonio Morgado, Jordi Planes, Alexey Ignatiev, and Joao Marques-Silva. Distance-restricted explanations: Theoretical underpinnings efficient implementation, 2024.

[26] Richard Dazeley, Peter Vamplew, Cameron Foale, Charlotte Young, Sunil Aryal, and Francisco Cruz. Levels of explainable artificial intelligence for human-aligned conversational explanations, 2021.

[27] Lukas Klein, Mennatallah El-Assady, and Paul F. Jäger. From correlation to causation: Formalizing interpretable machine learning as a statistical process, 2022.

[28] Pedro C. Neto, Tiago Gonçalves, João Ribeiro Pinto, Wilson Silva, Ana F. Sequeira, Arun Ross, and Jaime S. Cardoso. Causality-inspired taxonomy for explainable artificial intelligence, 2024.

[29] Ahmed M Salih. Explainable artificial intelligence and multicollinearity : A mini review of current approaches, 2024.

[30] Zhong Qiu Lin, Mohammad Javad Shafiee, Stanislav Bochkarev, Michael St. Jules, Xiao Yu Wang, and Alexander Wong. Do explanations reflect decisions? a machine-centric strategy to quantify the performance of explainability algorithms, 2019.

[31] Melkamu Mersha, Khang Lam, Joseph Wood, Ali AlShami, and Jugal Kalita. Explainable artificial intelligence: A survey of needs, techniques, applications, and future direction, 2025.

[32] Julie Gerlings, Arisa Shollo, and Ioanna Constantiou. Reviewing the need for explainable artificial intelligence (xai), 2021.

[33] AKM Bahalul Haque, A. K. M. Najmul Islam, and Patrick Mikalef. Explainable artificial intelligence (xai) from a user perspective- a synthesis of prior literature and problematizing avenues for future research, 2022.

[34] Ahmed Salih, Ilaria Boscolo Galazzo, Zahra Raisi-Estabragh, Steffen E. Petersen, Gloria Menegaz, and Petia Radeva. Characterizing the contribution of dependent features in xai methods, 2023.

[35] Md Montaser Hamid, Fatima Moussaoui, Jimena Noa Guevara, Andrew Anderson, Puja Agarwal, Jonathan Dodge, and Margaret Burnett. Inclusive design of ai's explanations: Just for those previously left out, or for everyone?, 2024.

[36] Samanta Knapič, Avleen Malhi, Rohit Saluja, and Kary Främling. Explainable artificial intelligence for human decision-support system in medical domain, 2021.

[37] Andrea Apicella, Luca Di Lorenzo, Francesco Isgrò, Andrea Pollastro, and Roberto Prevete. Strategies to exploit xai to improve classification systems, 2023.

[38] Behnaz Jamasb, Reza Akbari, and Seyed Raouf Khayami. On the applicability of explainable artificial intelligence for software requirement analysis, 2023.

[39] Yuyang Gao, Siyi Gu, Junji Jiang, Sungsoo Ray Hong, Dazhou Yu, and Liang Zhao. Going beyond xai: A systematic survey for explanation-guided learning, 2022.

[40] Ahmad Hussein, Mukesh Prasad, and Ali Braytee. Explainable ai methods for multi-omics analysis: A survey, 2024.

[41] Giulia Vilone and Luca Longo. Explainable artificial intelligence: a systematic review. *arXiv preprint arXiv:2006.00093*, 2020.

[42] Mir Riyanul Islam, Mobyen Uddin Ahmed, Shaibal Barua, and Shahina Begum. A systematic review of explainable artificial intelligence in terms of different application domains and tasks. *Applied Sciences*, 12(3):1353, 2022.

[43] Wojciech Samek and Klaus-Robert Müller. Towards explainable artificial intelligence, 2019.

[44] Samita Bai, Sidra Nasir, Rizwan Ahmed Khan, Sheeraz Arif, Alexandre Meyer, and Hubert Konik. Breast cancer diagnosis: A comprehensive exploration of explainable artificial intelligence (xai) techniques, 2024.

[45] Shideh Shams Amiri, Rosina O. Weber, Prateek Goel, Owen Brooks, Archer Gandley, Brian Kitchell, and Aaron Zehm. Data representing ground-truth explanations to evaluate xai methods, 2020.

[46] Jürgen Dieber and Sabrina Kirrane. Why model why? assessing the strengths and limitations of lime, 2020.

[47] Adam Perrett. Emergent explainability: Adding a causal chain to neural network inference, 2024.

[48] Weina Jin, Jianyu Fan, Diane Gromala, Philippe Pasquier, Xiaoxiao Li, and Ghassan Hamarneh. Transcending xai algorithm boundaries through end-user-inspired design, 2023.

[49] Laura State, Hadrien Salat, Stefania Rubrichi, and Zbigniew Smoreda. Explainability in practice: Estimating electrification rates from mobile phone data in senegal, 2023.

[50] Sédrick Stassin, Alexandre Englebert, Géraldin Nanfack, Julien Albert, Nassim Versbraegen, Gilles Peiffer, Miriam Doh, Nicolas Riche, Benoît Frenay, and Christophe De Vleeschouwer. An experimental investigation into the evaluation of explainability methods, 2023.

[51] Elvio G. Amparore, Alan Perotti, and Paolo Bajardi. To trust or not to trust an explanation: using leaf to evaluate local linear xai methods, 2021.

[52] Hubert Baniecki, Bartlomiej Sobieski, Patryk Szatkowski, Przemyslaw Bombinski, and Przemyslaw Biecek. Interpretable machine learning for time-to-event prediction in medicine and healthcare, 2024.

[53] José Ribeiro, Lucas Cardoso, Vitor Santos, Eduardo Carvalho, Níkolas Carneiro, and Ronnie Alves. How reliable and stable are explanations of xai methods?, 2024.

[54] Ugochukwu Orji and Elochukwu Ukwandu. Machine learning for an explainable cost prediction of medical insurance, 2023.

[55] Tirtha Chanda, Katja Hauser, Sarah Hobelsberger, Tabea-Clara Bucher, Carina Nogueira Garcia, Christoph Wies, Harald Kittler, Philipp Tschandl, Cristian Navarrete-Dechent, Sebastian Podlipnik, Emmanouil Chousakos, Iva Crnaric, Jovana Majstorovic, Linda Alhajwan, Tanya Foreman, Sandra Peternel, Sergei Sarap, İrem Özdemir, Raymond L. Barnhill, Mar Llamas Velasco, Gabriela Poch, Sören Korsing, Wiebke Sondermann, Frank Friedrich Gellrich, Markus V. Heppt, Michael Erdmann, Sebastian Haferkamp, Konstantin Drexler, Matthias Goebeler, Bastian Schilling, Jochen S. Utikal, Kamran Ghoreschi, Stefan Fröhling, Eva Krieghoff-Henning, and Titus J. Brinker. Dermatologist-like explainable ai enhances trust and confidence in diagnosing melanoma, 2023.

[56] Truong Thanh Hung Nguyen, Van Binh Truong, Vo Thanh Khang Nguyen, Quoc Hung Cao, and Quoc Khanh Nguyen. Towards trust of explainable ai in thyroid nodule diagnosis, 2023.

[57] Maryam Ahmed, Tooba Bibi, Rizwan Ahmed Khan, and Sidra Nasir. Enhancing breast cancer diagnosis in mammography: Evaluation and integration of convolutional neural networks and explainable ai, 2024.

[58] Amirehsan Ghasemi, Soheil Hashtarkhani, David L Schwartz, and Arash Shaban-Nejad. Explainable artificial intelligence in breast cancer detection and risk prediction: A systematic scoping review, 2024.

[59] Clive Gomes, Lalitha Natraj, Shijun Liu, and Anushka Datta. A survey of explainable ai and proposal for a discipline of explanation engineering, 2023.

[60] AKM Bahalul Haque, A. K. M. Najmul Islam, and Patrick Mikalef. Notion of explainable artificial intelligence – an empirical investigation from a users perspective, 2023.

[61] Neo Christopher Chung, Hongkyou Chung, Hearim Lee, Lennart Brocki, Hongbeom Chung, and George Dyer. False sense of security in explainable artificial intelligence (xai), 2024.

[62] Simon Hudson and Matija Franklin. Science communications for explainable artificial intelligence, 2023.

[63] Isaac Ronald Ward, Ling Wang, Juan lu, Mohammed Bennamoun, Girish Dwivedi, and Frank M Sanfilippo. Explainable artificial intelligence for pharmacovigilance: What features are important when predicting adverse outcomes?, 2021.

[64] Aida Brankovic, David Cook, Jessica Rahman, Wenjie Huang, and Sankalp Khanna. Evaluation of popular xai applied to clinical prediction models: Can they be trusted?, 2023.

[65] Ambreen Hanif. Towards explainable artificial intelligence in banking and financial services, 2021.

[66] Feini Huang, Shijie Jiang, Lu Li, Yongkun Zhang, Ye Zhang, Ruqing Zhang, Qingliang Li, Danxi Li, Wei Shangguan, and Yongjiu Dai. Applications of explainable artificial intelligence in earth system science, 2024.

[67] Timo Speith and Markus Langer. A new perspective on evaluation methods for explainable artificial intelligence (xai), 2023.

[68] Leander Weber, Sebastian Lapuschkin, Alexander Binder, and Wojciech Samek. Beyond explaining: Opportunities and challenges of xai-based model improvement, 2022.

[69] Tobias Labarta, Elizaveta Kulicheva, Ronja Froelian, Christian Geißler, Xenia Melman, and Julian von Klitzing. Study on the helpfulness of explainable artificial intelligence, 2024.

[70] Leilani H. Gilpin, Andrew R. Paley, Mohammed A. Alam, Sarah Spurlock, and Kristian J. Hammond. "explanation" is not a technical term: The problem of ambiguity in xai, 2022.

[71] Max Schemmer, Niklas Kühl, and Gerhard Satzger. Intelligent decision assistance versus automated decision-making: Enhancing knowledge work through explainable artificial intelligence, 2021.

[72] Adrien Bennetot, Ivan Donadello, Ayoub El Qadi, Mauro Dragoni, Thomas Frossard, Benedikt Wagner, Anna Saranti, Silvia Tulli, Maria Trocan, Raja Chatila, Andreas Holzinger, Artur d'Avila Garcez, and Natalia Díaz-Rodríguez. A practical guide on explainable ai techniques applied on biomedical use case applications, 2022.

[73] Sven Nomm. Towards the linear algebra based taxonomy of xai explanations, 2023.

[74] Andrés Páez. The pragmatic turn in explainable artificial intelligence (xai), 2020.

[75] Barnaby Crook, Maximilian Schlüter, and Timo Speith. Revisiting the performance-explainability trade-off in explainable artificial intelligence (xai), 2023.

[76] Behnam Mohammadi, Nikhil Malik, Tim Derdenger, and Kannan Srinivasan. Regulating explainable artificial intelligence (xai) may harm consumers, 2024.

[77] Jacob Dineen, Don Kridel, Daniel Dolk, and David Castillo. Unified explanations in machine learning models: A perturbation approach, 2024.

[78] Matija Franklin. The influence of explainable artificial intelligence: Nudging behaviour or boosting capability?, 2022.

[79] Jamie Duell, Monika Seisenberger, Gert Aarts, Shangming Zhou, and Xiuyi Fan. Towards a shapley value graph framework for medical peer-influence, 2022.

[80] Weina Jin, Xiaoxiao Li, and Ghassan Hamarneh. Why is plausibility surprisingly problematic as an xai criterion?, 2024.

[81] Xueting Huang, Zhibo Zhang, Fusen Guo, Xianghao Wang, Kun Chi, and Kexin Wu. Research on older adults' interaction with e-health interface based on explainable artificial intelligence, 2024.

[82] Matteo Rizzo, Alberto Veneri, Andrea Albarelli, Claudio Lucchese, Marco Nobile, and Cristina Conati. A theoretical framework for ai models explainability with application in biomedicine, 2023.

Disclaimer:

SurveyX is an AI-powered system designed to automate the generation of surveys. While it aims to produce high-quality, coherent, and comprehensive surveys with accurate citations, the final output is derived from the AI's synthesis of pre-processed materials, which may contain limitations or inaccuracies. As such, the generated content should not be used for academic publication or formal submissions and must be independently reviewed and verified. The developers of SurveyX do not assume responsibility for any errors or consequences arising from the use of the generated surveys.