EXPLAINABLE ARTIFICIAL INTELLIGENCE (XAI): BRIDGING TRANSPARENCY AND TRUST IN MACHINE LEARNING SYSTEMS

Noman Javed[*1], Noshad Ali[2], Kamal Khan[3], Imtiaz Kamal[4], Khalid Ali[5],
Satyadhar Joshi[6], Rabia Altaf Kalhoro[7], Zohra Naim[8]

[*1]*University of West Scotland London Campus*
[2]*Faculty of engineering science and technology, department of computer science, lasbela university of Agriculture, Water, and Marine Sciences Uthal.*
[3]*University of Makran*
[4]*University Of Turbat*
[5]*Department of Computer Science, University college of Dera Murad Jamali, Lasbela University of Agriculture, Water and Marine Science Uthal, Balochistan, Pakistan.*
[6]*Information Technology Department, MSIT Alumnus, Touro College, New York, NY 10010, USA.*
[7,8]*Sindh Madressatul Islam university*

[*1]noman19304@gmail.com, [2]noshadali_cs@outlook.com, [3]kamalkhan@uomp.edu.pk,
[4]imtiazkamal181@gmail.com, [5]khalid.ali@ucdmj.luawms.edu.pk, [6]sjoshi@student.touro.edu,
[7]rabiakalhoroaltaf21@gmail.com, [8]zohranaim1996@gmail.com

DOI: https://doi.org/10.5281/zenodo.17719522

Keywords
Explainable Artificial Intelligence (XAI), Transparency, Trust, Interpretability, Model-Agnostic Methods, LIME, SHAP, Ethical AI, Healthcare Applications, Finance AI, Autonomous Systems, Black-Box Models, Human-Centered Design

Article History
Received: 06 October 2025
Accepted: 15 November 2025
Published: 26 November 2025

Corresponding Author: *
Noman Javed

Abstract

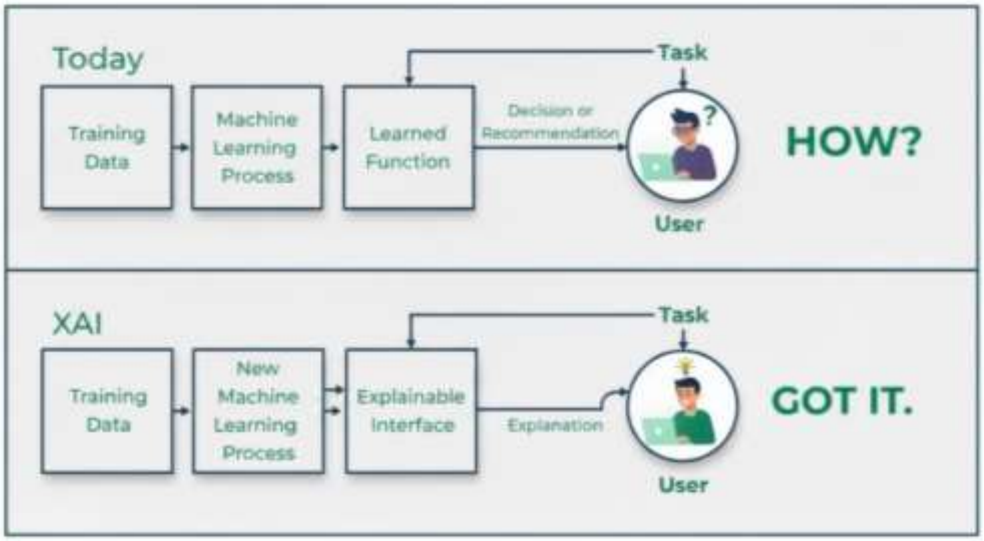
Explainable Artificial Intelligence (XAI) addresses the opacity of complex machine learning models by enhancing transparency, interpretability, and trust in AI systems. This paper explores the fundamental principles of XAI, including transparency, accountability, and fairness, while delineating the need for explainability in high-stakes domains such as healthcare, finance, and autonomous systems. It categorizes XAI methods into model-specific and model-agnostic techniques, such as LIME and SHAP, and examines their real-world applications for improving decision-making, regulatory compliance, and ethical AI deployment. Challenges like balancing accuracy with interpretability, user-specific explanations, and standardization of metrics are discussed, alongside future directions emphasizing human-centered design. Through a comprehensive review, the paper underscores XAI's role in fostering responsible AI adoption, mitigating biases, and bridging the gap between advanced AI performance and human understanding.

INTRODUCTION

Explainable Artificial Intelligence (XAI) involves AI methods and techniques aimed at improving the interpretability and understandability of machine learning outcomes. The foremost aim of XAI is to explain the inner workings of an AI system and how it arrives at a decision, thereby offering a bit of transparency and explainability, especially regarding the intricate deep learning systems where XAI is most needed. (Pavlidis, 2024). XAI provides an explanation of the various AI processes which helps users and stakeholders to trust it, thereby promoting acceptance and responsible engagement of the AI systems (Morato De Andrade & Sousa Alves, 2024). XAI illuminates the decision-making processes of automated systems, thus, fulfilling the requirements of regulations that demand explainability (Wolf & Ringland, 2020). Understanding model predictions helps improve the design and performance of machine learning systems by providing the means necessary to remove, biases, and errors (Ali et al., 2023). Particularly in high-stakes domains, such as AI-assisted healthcare, AI-assisted finance, and AI-assisted legal decision making, XAI directly enhances the users' ability to rationally interpret the AI output and justify their decisions (Ali et al., 2023; Górski & Ramakrishna, 2021). In the realm of Ethical AI Development, the contribution of XAI (Explainable Artificial Intelligence) is fundamental to the construction of equitable and non-discriminatory AI systems. This is essential for preventing inequality and fostering ethically positive implications (Morato De Andrade & Sousa Alves, 2024). Like XAI, AI techniques transform systems to enhance the transparency of trust built during the processes of decision making. AI systems, most of which are still described and classified as "black-box" models, have become significantly less opaque to human users since the introduction of "explainable AI" technologies. (Wolf & Ringland, 2020).

One of the most direct and common approaches to transparency built on trust is the interpretability of AI systems. These models rely on approaches such as saliency maps, attention mechanisms, rule-based explanations, and other model-agnostic frameworks to AI interpretability (Thalpage, 2023). These approaches culminate in user-facing aids that provide decision rationales, critical for augmenting trust in autonomous systems (Mathew et al., 2025). On top of that, XAI promotes the ethical use of AI by making sure AI results meet the user's needs, expectations, and regulations. In the finance sector, XAI affords institutions the opportunity to fulfill compliance by delivering transparent and accountable AI-enabled decisions, thus trust by consumers, and strengthening the regulation playground (Anang et al., 2024). Concerning trust, XAI builds confidence through the comprehension and explication of diverse usage. In the clinical setting, for example, XAI helps clinicians to comprehend AI-enabled medical diagnoses, which builds trust and helps in the assimilation of AI in the medical practice (Chanda et al., 2024).To gain confidence, XAI argues for the education of users to gain meaningful transparency. Users need to understand AI technology and the architecture of decisions made by these systems to enhance confidence and the informed interaction of users (Olateju et al., 2024). XAI helps to understand and counter the negative effects of bias in AI systems, particularly in high-stake areas like healthcare and criminal justice, where ethics and fairness are crucial (R, 2024). XAI promotes trust through transparency by explaining systems and decisions made, which helps to build the accountability and fairness of the provided AI systems. As discussed earlier, the bridge between transparency and trust in AI systems hinges on XAI. It has the potential of advancing ethical and responsible adoption of AI technologies in many industries, primarily because it enhances the clarity and interpretability of AI systems (Belghachi, 2023).

Figure 1.1 Evolution from Opaque Machine Learning to Transparent XAI: Enhancing User Comprehension in Decision Processes



2. The Requirement of Explainability

In the AI field, a system of arbitrary complexity can achieve a given task to sufficient accuracy and resolution (e.g., healthcare, finance, and security). Nevertheless, systems built using machine learning and deep learning techniques can produce excellent results at a complexity level beyond the comprehension of even the system's builders (Islam et al., 2022; Marey et al., 2024; Shahrzad et al., 2025). The "black box" label raises the issue of a system's decision process being potentially understand. This lack of understanding stifles trust, which is of particular importance in the use of AI in critical settings, such as healthcare and forensic settings. Ethical and legal questions of accountability, bias, non-maleficence, and beneficence, among others, are entailed in opaque systems. (Adeniran et al., 2024; Marey et al., 2024). The approach taken to mitigate the black box problem resulted in the creation of Explainable AI (XAI) solutions. Part of the goal of the XAI approach is to create models that both excel at tasks and help understand the rationale behind the decisions taken, thereby improving interpretability and transparency. These techniques provide justification to users, helping to trust the AI systems, which is vital in gaining acceptance and ethical deployment (Fleisher, 2022; Islam et al., 2022). Being able to explain the decisions made by Artificial Intelligence systems aids in fostering ethical, fair, and trustworthy outcomes. Because of the direct correlation between the explainability of AI systems and achieving ethical AI, fairness, and trust, ubiquitous transparency is vital to the ethical deployment of AI. (Alonso 2020; Pavlidis 2024).

First, explainability facilitates the ethical use of AI by clarifying the decision-making processes that AI systems automate. Stakeholders can comprehend the reasoning behind a decision and how it was made to evaluate whether it meets ethical and societal values (Bulut & Beiting-Parrish, 2024; Owolabi et al., 2024). Transparency, therefore, helps stakeholders, including designers, occupants, and regulators, discern the inequalities, and unethical outcomes that the system may generate, thereby revealing the biases hidden in the system. On the issue of fairness, explainability helps pinpoint biases in AI systems and suggest ways to alleviate them. These biases may permeate algorithms, resulting in unfair, discriminatory treatment of certain populations. To the extent that explainability provides insight into the decision-making processes of AI systems, stakeholders will be in a better position to identify the biases in a given context and implement changes to mitigate the resulting discrimination

(Gupta, 2023). Explainability also facilitates the formulation of equity-optimizing outcomes and significantly impacts the incorporation of fairness and equity in AI operations (Pham, 2025). Explainability is a prerequisite for the trustworthiness of AI systems, which encompasses reliability, accountability, and transparency. Users will only trust AI systems if they can discern the logic behind decisions and the systems can produce a coherent account of that logic. This trust is necessary to foster confidence and facilitate the incorporation of AI into a selected industry, be it healthcare, finance, or autonomous systems (Chamola et al., 2023; Udegbe et al., 2024). This explains the growing incorporation of explainability in recent AI models (Kaur et al., 2022; Liu et al., 2022).

3. Core Principles of XAI

Differentiating between 'explainability' and 'interpretability' is a topic of ongoing debate in the field of artificial intelligence (AI). While both concepts relate to some form of transparency in AI systems, their meanings and implications do not align. (Alonso 2020; Pavlidis 2024).

As opposed to explainability, interpretability is concerned with how well a human being understands the causal links between the components of a model and how their outputs interact with one another. Most fundamentally, it focuses on how the reasoning of an AI model can be made understandable to a human, preferably in a manner whereby the logic of each step the AI undertakes can be traced. In the case of AI systems used in the healthcare sector, especially AI-supported clinical healthcare decision systems, interpretability is needed to build trust (Xu et al., 2023). While there is a functional overlap between explainability and interpretability, explainability focuses on a broader scope, and includes articulating and justifying a model's outputs in a way that is understandable to users. Explainability is concerned with how the reasoning behind a model's decisions can be described to users. Important factors to consider in explainability include the objectives of the explanation, the situation in which the explanation is provided, and the users in question (Ferreira & Monteiro, 2020). In accountable domains like health care and finance, it is necessary to offer simplified explanations of the model, as complexity can obscure understanding (Pavlidis, 2024).

The "black box" problem in machine learning, especially with deep learning, is highly performant but lacks clarity. There is an ongoing imbalance in the transparency of the explainability performance of AI systems (Raz et al., 2024). Enhancing the transparency and explainability of AI model outputs is essential for trustworthiness and the broader use of such systems (Holzinger et al., 2019). Explainable Artificial Intelligence (XAI) is defined by the principles of transparency, accountability, and fairness, which are the foundations of the inclusion of AI systems in critical sectors such as healthcare, finance, and law. (R, 2024). Within XAI, transparency means making AI systems comprehensible to people by explaining how algorithms arrive at conclusions and breaking down the "black box" opacity of the AI model, allowing users to see and understand the processes behind the AI's decisions. This fosters trust, particularly in high-stakes domains such as medicine and law, where decisions can profoundly affect lives (Karim et al. 2023; Thalpage 2023). Within the context of AI, accountability involves making AI developers and operators responsible for the outcomes of their AI systems. This means that the decision-making processes must be not only clear and exposed, but also traceable, so that biases and mistakes can be found. This principle makes certain that AI systems are applied ethically, and that stakeholders can be sure the system operators will be responsible for any negative consequences of the AI decisions (Alonso 2020; Pavlidis 2024). Fairness pertains to XAI in that explainable AI must not be biased and must not discriminate against any users. This means that equal treatment to all users must be engineered by reducing algorithmic bias. This becomes especially significant when evaluating AI systems used for life-impacting decisions, such as credit scoring and criminal justice, where the principles concerning fairness as an obligation must apply (Karim et al., 2023; Thalpage, 2023).

Table 3.1 Core Principles of XAI

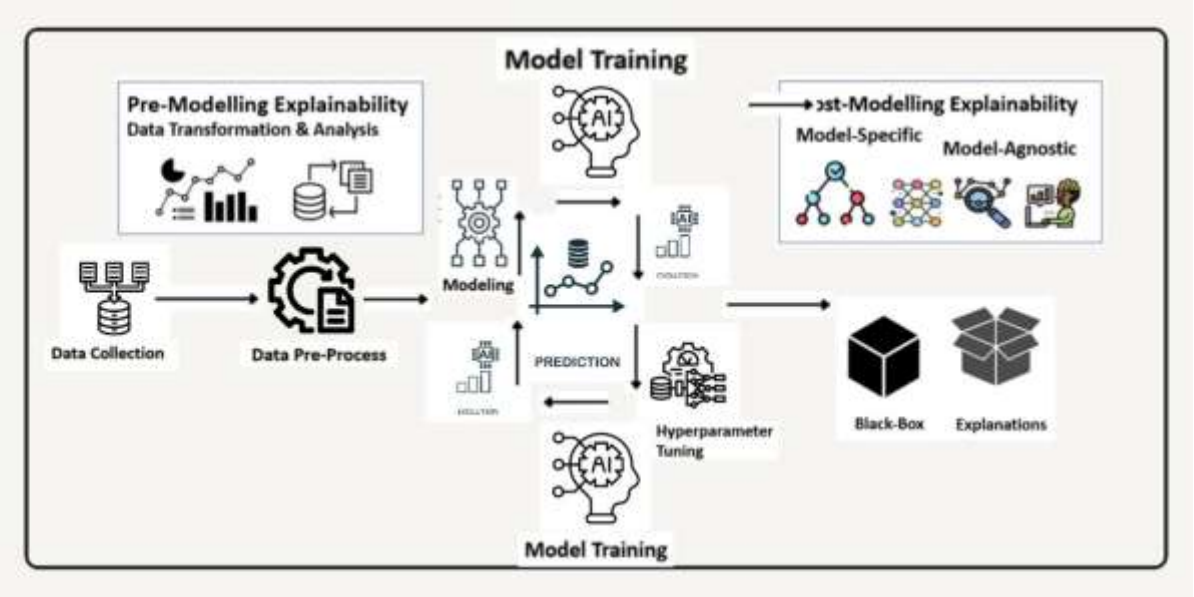
Principle	Description	Citation(s)
Transparency	Makes AI processes comprehensible by explaining algorithms and breaking down opacity.	(Karim et al., 2023; Thalpage, 2023)
Accountability	Ensures developers are responsible for outcomes, with traceable processes to identify biases.	(Alonso, 2020; Pavlidis, 2024)
Fairness	Reduces algorithmic bias to ensure equal treatment, especially in life-impacting decisions.	(Karim et al., 2023; Thalpage, 2023)
Interpretability	Focuses on understanding causal links in models for traceability.	(Xu et al., 2023)
Explainability	Articulates model outputs in user-understandable ways, considering context and objectives.	(Ferreira & Monteiro, 2020; Pavlidis, 2024)

4. Methods and Techniques

As part of Explainable Artificial Intelligence (XAI), there are primarily two broad categories of techniques used: model-specific and model-agnostic approaches. (Ehsan et al., 2021) Model-Specific Methods: A model-specific approach focuses on the explanation of an artificial intelligence model specifically. These techniques utilize the built reflective elements of the model and explain how the particular model arrived at its outcomes. For example, the Grad-Cam method serves as an explanation for convolutional neural networks so that users can see which parts of the image are most responsible for the network's decision, thus, explaining the reasoning of the model contextually within the model architecture (Amin et al., 2023). Similarly, interpretable deep learning models, attention-based models, and other incorporated systems were designed so that their outputs are more understandable in the reasoning process (Thalpage, 2023). Model-Agnostic Methods: Unlike other approaches, model-agnostic methods can be used with any type of machine learning model in any architecture. These methods produce explanations of the predictions that are independent of the model, thus allowing for diverse uses across different models and applications. Well-known model-agnostic methods are the following: (Pavlidis, 2024).

The model LIME explains the predictions of a model by locally approximating the model with simpler, interpretable models, thus aiding in the understanding of individual predictions (Zahoor et al., 2024). SHAP explains predictions by calculating, through cooperative game theory, the exact contribution of each feature in a prediction, thus allowing fair attribution of prediction responsibility to each feature (Belghachi, 2023). These plots and visualizations illustrate the relationship between the predictors and the predicted value, thus aiding users in understanding how changes in predictor values influence the predictions (Wang, 2024). Each method has certain strengths and weaknesses in interpretability and computational resources, and their use is mostly determined by the context and nature of the data (Wang, 2024). The decision between model-specific and model-agnostic approaches is, in most cases, determined by the application in use and the complexity of the model. LIME (Local Interpretable Model-Agnostic Explanations) and SHAP (SHapley Additive exPlanations) are powerful tools designed to aid transparency in machine learning by enhancing the interpretability and understandability of models' output and the underlying logic of the models' predictions. These tools are part of Explainable AI (XAI), which attempts to eliminate the "black box" characteristics of advanced algorithms and offer better understandings of the predictive rationales of the models. LIME is designed to function as a model-agnostic explainability framework which emphasizes its widespread applicability and use across various machine learning models. It works by perturbing the input variables and tracking the output alterations to approximate a model's decision boundary locally around the input. This process is done by simulating a simpler and interpretable model that approximates the function of the complex model around the specific data point of interest (Korade, 2024). LIME contributes to explainability by offering locally relevant rationales and assisting users in assessing the impact of specific variables in a prediction (Vimbi et al., 2024).

Figure 4.1 Integrating Explainability Throughout the Machine Learning Workflow: Pre- and Post-Modeling Strategies



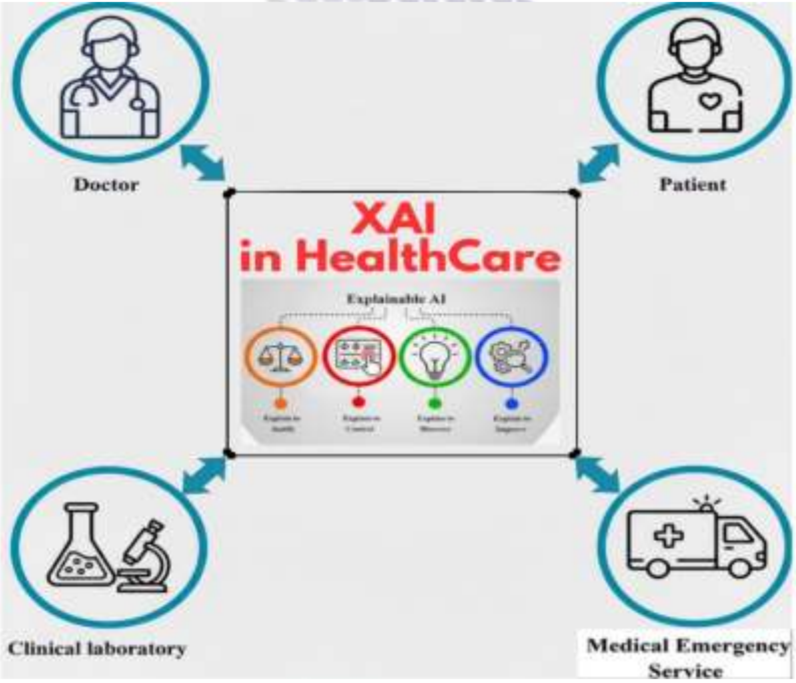
Conversely, SHAP operates on robust theoretical principles, utilizing Shapley values from cooperative game theory. It equitably allocates a prediction among the input features by assessing different feature combinations and determining how each feature contributes to the prediction. Consistently, SHAP delivers locally accurate rationales that explain the difference between the output value the model averages and the actual output for the particular instance (Hendriks & al., 2020; Korade, 2024). This approach reinforces the notion that the attributions explain the difference between the actual and expected outputs of the model, providing a concise explanation that holds true across various model types and applications (Hasan, 2024). LIME, SHAP is used in numerous disciplines, including healthcare, cybersecurity, and financial forecasting, where the rationale of AI models must be understood to uphold accuracy, fairness, and trustworthiness (Chen et al., 2023; Gaspar et al., 2024). These methods are pivotal in revealing model biases, determining the important features that drive model predictions, and providing guidance to end users that enhances the model's trustworthiness and overall performance (Khan et al., 2024; Ma et al., 2023). Each of the two techniques has distinct strengths and weaknesses despite the unique insights they each provide. LIME allows for rapid local interpretability because of its simplicity and flexibility, although perturbed instances may lead to explanations that lack stability. On the other hand, although SHAP has a greater computational demand, it is still the better approach to achieving a coherent and consistent global perspective given its definitive articulation of a feature's contribution to the prediction (Salih et al., 2024). In summary, each of the techniques provides essential contributions to transparency and accountability of machine learning systems to model audits, and ultimately, user trust (Hasan, 2024; Korade, 2024).

5. Real-World Applications

Explainable Artificial Intelligence (XAI) improves transparency, trust, and accountability in the application of AI tools, and is a game changer in sectors like healthcare, finance, and autonomous systems. (Panda & Mahanta, 2024).XAI provides meaningful healthcare transparency and AI model interpretability in the prediction and diagnostics of diseases. Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) methods, designed to interpret model predictions, are common in the medical field (Alkhanbouli et al., 2025). These approaches help alleviate the 'black-box' concern of AI, fostering trust with healthcare providers by explaining details of a prediction, much like a case-analytic prediction (Metta et al., 2024).

XAI facilitates healthcare AI systems in decision making to support functions that include diagnostics of diseases and treatment recommendations tailored to the individual, and encourages ethical adherence (Adeniran et al., 2024; Zhang et al., 2022). Within the financial sector, XAI addresses the conflicting priorities of innovation and regulatory adherence. XAI contributes to the transparency of AI applications in fraud detection, credit scoring, and algorithmic trading, which fulfills regulatory demands for transparency and accountability (Anang et al., 2024). This is vital in upholding consumer trust, while also providing room for innovation in financial technologies. Additionally, XAI improves risk management and portfolio optimization by offering explainable decisions, which facilitate risk assessments and regulatory adherence (Černevičienė & Kabašinskas, 2024; Weber et al., 2023).

Figure 5.1 Role of Explainable AI in Healthcare Ecosystems: Facilitating Justification, Control, Discovery, and Improvement Across Stakeholders



In autonomous systems and more specifically autonomous vehicles, XAI is vital for the justification of decisions which fosters safety and builds public trust in the technology. XAI fosters safety and trust needed for the commercialization of autonomous vehicles by detailing the systems used to navigate and make decisions in real time (Madhav & Tyagi, 2022). This transparency fosters trust while the systems integrate with human-driven vehicles (Veitch & Alsos, 2021). The use of XAI in these important fields demonstrates the increasing focus on transparent and accountable AI systems which are understandable and interpretable by human beings. While there are still challenges to overcome such as the trade-off between the complexity of the model and its interpretability as well as ethical concerns around bias, XAI still improves the functionality and trust of AI systems in health care, finance, and autonomous systems (Belghachi, 2023). XAI offers explanations as to how AI models make decisions, which is necessary for accountability. AI systems primarily function as 'black boxes' with little to no explanations being offered post decisions. With explanations, users' ability to comprehend the rationale behind a decision is significantly improved, suggesting systems can be held accountable (Adeniran et al., 2024; Pawar et al., 2020).

AI systems have the potential to positively impact user confidence, but only within the context of systems that provide explanations. Systems that offer 'simplified' visualizations of 'predictions' enhance user confidence, which trust is built upon (Hudon et al., 2021). The literature shows mixed findings on the direct impact of XAI on decision-making. XAI trust is built when users are offered explanations to AI predictions, thus increasing their decision-making abilities. Subsequently, they can identify reliable and faulty decisions based on the explanations given (Alufaisan et al., 2021). XAI provides explain ability, enabling stakeholders to evaluate and determine if AI-driven decisions are ethical. Demonstrating ethical AI decisions will help build trust in AI systems. (Adeniran et al, 2024; Upadhyay et al, 2023). Advancing functional transparency contributes to the development of systems that fulfill societal expectations. This cultivates trust in AI systems and acceptance in verticals. (Hosaïn et al, 2023).

Table 5.1 Real-World Applications of XAI

Sector	Application	Benefits	Citation(s)
Healthcare	Disease Prediction and Diagnostics	Enhances transparency, builds clinician trust, and supports ethical decisions.	(Alkhanbouli et al., 2025; Metta et al., 2024; Adeniran et al., 2024; Zhang et al., 2022)
Finance	Fraud Detection, Credit Scoring	Meets regulatory demands, improves risk management, and fosters consumer trust.	(Anang et al., 2024; Černevičienė & Kabašinskas, 2024; Weber et al., 2023)
Autonomous Systems	Vehicle Navigation and Decisions	Justifies real-time decisions for safety and public trust in commercialization.	(Madhav & Tyagi, 2022; Veitch & Alsos, 2021)

6. Challenges and Limitations

One of the key challenges in AI technology is the balance between model accuracy and interpretability. This becomes increasingly important with the rise of AI in the healthcare, financial, and data center management industries where trust and transparency are key components. (R, 2024). The more accurate an AI model is, the more convoluted it is, which

makes it more difficult to understand. This is the case with deep learning models, which are accurate, but opaque and difficult to understand, leading to a 'black box' situation. This makes the challenge of AI being effective, understandable, and feasible trustworthy a more daunting task (Ennab & Mcheick, 2024; Mathew et al., 2025). The uncertainty of AI system outputs can undermine stakeholders' and users' confidence, trust, and reliance on the AI system. This is particularly the case in healthcare, where AI-assisted decisions can lead to life-threatening and harmful decisions, as the AI outputs cannot be explained. The same situation is present in the financial services sector, where decision-makers cannot trust AI systems as the decision rationales are not accessible (Ennab & Mcheick, 2022;

Gebreyesus et al., 2024). AI system transparency is legally mandated and is necessary to ethically align with the intended purpose of the system. However, the demands of transparency conflict with the intricacies and complexities intertwined in systems of high accuracy, which requires a rigorous focus and energy towards building interpretability (without highly sacrificing accuracy) (R, 2024; Raz et al., 2024).

These methods focus on building understanding and trust by analyzing feature importance, understanding the reasoning behind decisions, and finding the balance needed to implement them without compromising model performance poses a technical obstacle (Gebreyesus et al., 2024; Sabharwal et al., 2024). The tension between the prioritization of either accuracy or interpretability is an interdisciplinary challenge, incorporating technical, ethical, and domain-specific perspectives. This is a key contributor to the discord surrounding the balance placed on these aspects and the ensuing difficulty in developing AI systems that attain several desired goals (Raz et al., 2024). The primary causes stem from the diversity of the applications, the intricacies of the models, and the disparate needs of the stakeholders involved. Different user types have varying degrees of complexity concerning their explanations. While a data scientist may seek a deeper understanding of a machine learning model's components, a casual user might simply want a model's function described in layman's terms (Pavlidis, 2024). The need to develop personalized communication for the varied user types makes the quest for universal metrics more intricate. In the case of many AI technologies, and particularly those based on deep learning, the algorithms operate in an opaque and complex manner (Emmert-Streib et al., 2020). Given the diverse complexities and configurations of machine learning models, it is inevitable that a universal metric may not be adequate nor applicable to varying algorithms. Concerning the assess ability of the diversity in explainability techniques, which include but are not limited to LIME, SHAP and counterfactual explanations, each technique presents unique aspects, advantages and limitations that make it incredibly challenging to formulate a universal metric that quantifies the explanatory value and profundity of disparate techniques (Panda & Mahanta, 2024).

Several domains, such as healthcare, finance and criminal justice, impose specific and distinct expectations on explainability (R, 2024). Differences in the type of explanations required by these domains will lead to the need variations in measurement, which will make it more difficult to standardize metrics. In some domains, explainability might not only be a technical challenge, but a legal and ethical one, which adds more complexity. The AI Act of the European Union underscores the importance of explainability, but the techniques and standards required still need to be delineated (Mathew et al., 2025). There is little agreement in the field about explaining evaluations. Some metrics focus on plausibility, the fidelity of the explanation, while others focus on interpretability, satisfaction, and diverging approaches in different studies (Geng, 2024). The explain ability of more complicated AI models poses a challenge. Certain explainability techniques may be appropriate to use on simpler models and then fail as the models become more complex, resulting in a challenge to developing standardized metrics to explain the variability (R, 2024).

Table 6.1 Challenges and Limitations in XAI

Challenge	Description	Citation(s)
Accuracy vs. Interpretability	High-accuracy models (e.g., deep learning) are opaque, complicating trust in critical sectors.	(Ennab & Mcheick, 2024; Mathew et al., 2025)
User-Specific Explanations	Varying needs across users (e.g., experts vs. laypeople) make universal metrics difficult.	(Pavlidis, 2024)
Model Complexity	Deep learning opacity hinders universal applicability of explainability techniques.	(Emmert-Streib et al., 2020)

Bias and Ethical Concerns	Identifying and mitigating biases in diverse domains requires domain-specific metrics.	(Panda & Mahanta, 2024; R, 2024)
Scalability	Techniques work on simple models but fail on complex ones, challenging standardization.	(R, 2024)
Evaluation Metrics	Lack of agreement on metrics like plausibility vs. fidelity across studies.	(Geng, 2024)

7. Future Directions

Human-centered design has the potential to improve the effectiveness of Explainable Artificial Intelligence (XAI). By aligning the AI systems to the needs of the users, interactions become more fruitful. A focus on human-centered design is the importance of confirmable practicality as interpretability is for real users and not only for the algorithm developers. (Hudon et al., 2021). Understanding AI and machine learning is the first step toward the collaboration of users like game designers and AI/ML tools (Zhu et al., 2018). Fulfilling user expectations is crucial for operationalizing XAI within a human-centered approach at the conceptual, methodological, and technical levels. This implies the development of constructive frameworks, assessment strategies, and design principles that support the development of AI systems that are not only transparent and understandable but also facilitate user trust and human-AI cooperation (Ehsan et al., 2021).

Human-centered explainable AI (HCXAI) is concerned with the user's operational environment and requirements. This approach is concerned with the actionability and significance of explanatory components. It advocates the value of situational awareness and the user's story in system transparency and accountability, helpful in AI-integrated decision systems to build trust, ensuring the user's explanations are functional (Ridley, 2024). Evaluating XAI from a human-centered perspective also involves determining what constitutes a meaningful explanation to a user. This entails understanding explanation quality in context, contribution to human-AI interaction, and effects on human-AI performance. Through the lens of human-centric assessments, researchers will be able to determine the explanation-value attributes and form a taxonomy on the evaluation of human-centered XAI (Kim et al., 2024). Finally, the more visualization and simplification methods reduce user cognitive load, the more they will feel confident in the AI system. Users will understand AI-driven predictions and decisions better, and trust will follow as the cognitive load becomes more manageable in the collaboration (Hudon et al., 2021).

References

Adeniran, A., William, P., & Onebunne, A. (2024). Explainable AI (XAI) in healthcare: Enhancing trust and transparency in critical decision-making. World Journal of Advanced Research and Reviews, 23(3), 2447–2658. https://doi.org/10.30574/wjarr.2024.23.3.2936

Ali, S., Akhlaq, F., Imran, A. S., Kastrati, Z., Daudpota, S. M., & Moosa, M. (2023). The enlightening role of explainable artificial intelligence in medical & healthcare domains: A systematic literature review. Computers in Biology and Medicine, 166, 107555. https://doi.org/10.1016/j.compbiomed.2023.107555

Alkhanbouli, R., Matar Abdulla Almadhaani, H., Alhosani, F., & Simsekler, M. C. E. (2025). The role of explainable artificial intelligence in disease prediction: a systematic literature review and future research directions. BMC Medical Informatics and Decision Making, 25(1). https://doi.org/10.1186/s12911-025-02944-6

Alonso, J. M. (2020). Teaching Explainable Artificial Intelligence to High School Students. International Journal of Computational Intelligence Systems, 13(1), 974.

https://doi.org/10.2991/ijcis.d.200715.003

Alufaisan, Y., Marusich, L. R., Kantarcioglu, M., Bakdash, J. Z., & Zhou, Y. (2021). Does Explainable Artificial Intelligence Improve Human Decision-Making? Proceedings of the AAAI Conference on Artificial Intelligence, 35(8), 6618–6626. https://doi.org/10.1609/aaai.v35i8.16819

Amin, A., Hasan, K., Ahmed, I., Islam, T., Zein-Sabatto, S., & Chimba, D. (2023). An Explainable AI Framework for Artificial Intelligence of Medical Things. 2097–2102. https://doi.org/10.1109/gcwkshps58843.2023.10464798

Anang, A., Arogundade, J., Sonubi, T., Akinbi, I., Nwafor, K., & Ajewumi, O. (2024). Explainable AI in financial technologies: Balancing innovation with regulatory compliance. International Journal of Science and Research Archive, 13(1), 1793–1806. https://doi.org/10.30574/ijsra.2024.13.1.1870

Belghachi, M. (2023). A Review on Explainable Artificial Intelligence Methods, Applications, and Challenges. Indonesian Journal of Electrical Engineering and Informatics (IJEEI), 11(4). https://doi.org/10.52549/ijeei.v11i4.5151

Bulut, O., & Beiting-Parrish, M. (2024). The Rise of Artificial Intelligence in Educational Measurement: Opportunities and Ethical Challenges. Chinese/English Journal of Educational Measurement and Evaluation, 5(3). https://doi.org/10.59863/miql7785

Černevičienė, J., & Kabašinskas, A. (2024). Explainable artificial intelligence (XAI) in finance: a systematic literature review. Artificial Intelligence Review, 57(8). https://doi.org/10.1007/s10462-024-10854-8

Chamola, V., Sikdar, B., Hassija, V., Dhingra, D., Sulthana, A. R., & Ghosh, D. (2023). A Review of Trustworthy and Explainable Artificial Intelligence (XAI). IEEE Access, 11, 78994–79015. https://doi.org/10.1109/access.2023.3294569

Chanda, T., Rasulova, G., Garzona-Navas, L., Welponer, T., Theofilogiannakou, P., Iglesias-Pena, N., Dragolov, M., Yanatma, I., Golle, L., Ghoreschi, K., Chousakos, E., Drexler, K., Schmitt, L., Schilling, B., Cenk, H., Peternel, S., Thiem, A., Peralta, R., Hobelsberger, S., … Ferhatosmanoğlu, A. (2024). Dermatologist-like explainable AI enhances trust and confidence in diagnosing melanoma. Nature Communications, 15(1). https://doi.org/10.1038/s41467-023-43095-4

Chen, Y., Calabrese, R., & Martin-Barragan, B. (2023). Interpretable machine learning for imbalanced credit scoring datasets. European Journal of Operational Research, 312(1), 357–372. https://doi.org/10.1016/j.ejor.2023.06.036

Ehsan, U., Wachter, S., Wintersberger, P., Riedl, M. O., Liao, Q. V., Mara, M., Riener, A., & Streit, M. (2021). Operationalizing Human-Centered Perspectives in Explainable AI. 31, 1–6. https://doi.org/10.1145/3411763.3441342

Emmert-Streib, F., Yli-Harja, O., & Dehmer, M. (2020). Explainable artificial intelligence and machine learning: A reality rooted perspective. WIREs Data Mining and Knowledge Discovery, 10(6). https://doi.org/10.1002/widm.1368

Ennab, M., & Mcheick, H. (2022). Designing an Interpretability-Based Model to Explain the Artificial Intelligence Algorithms in Healthcare. Diagnostics, 12(7), 1557. https://doi.org/10.3390/diagnostics12071557

Ennab, M., & Mcheick, H. (2024). Enhancing interpretability and accuracy of AI models in healthcare: a comprehensive review on challenges and future directions. Frontiers in Robotics and AI, 11. https://doi.org/10.3389/frobt.2024.1444763

Ferreira, J. J., & Monteiro, M. S. (2020). What Are People Doing About XAI User Experience? A Survey on AI Explainability Research and Practice (pp. 56–73). Springer. https://doi.org/10.1007/978-3-030-49760-6_4

Fleisher, W. (2022). Understanding, Idealization, and Explainable AI. Episteme, 19(4), 534–560. https://doi.org/10.1017/epi.2022.39

Gaspar, D., Silva, C., & Silva, P. (2024). Explainable AI for Intrusion Detection Systems: LIME and SHAP Applicability on Multi-Layer Perceptron. IEEE Access, 12, 30164–30175. https://doi.org/10.1109/access.2024.3368377

Gebreyesus, Y., Chinnici, M., Dalton, D., Chinnici, A., & De Chiara, D. (2024). AI for Automating Data Center Operations: Model Explainability in the Data Centre Context Using Shapley Additive Explanations (SHAP). Electronics, 13(9), 1628. https://doi.org/10.3390/electronics13091628

Geng, S. (2024). Analysis of the Different Statistical Metrics in Machine Learning. Highlights in Science, Engineering and Technology, 88, 350–356. https://doi.org/10.54097/jhq3tv19

Górski, Ł., & Ramakrishna, S. (2021). Explainable artificial intelligence, lawyer's perspective. 31, 60–68. https://doi.org/10.1145/3462757.3466145

Gupta, N. (2023). Artificial Intelligence Ethics and Fairness: A study to address bias and fairness issues in AI systems, and the ethical implications of AI applications. Revista Review Index Journal of Multidisciplinary, 3(2), 24–35. https://doi.org/10.31305/rrijm2023.v03.n02.004

Hasan, M. M. (2024). Understanding Model Predictions: A Comparative Analysis of SHAP and LIME on Various ML Algorithms. Journal of Scientific and Technological Research, 5(1), 17–26. https://doi.org/10.59738/jstr.v5i1.23(17-26).eaqr5800

Hendriks, M. P., Ten Teije, A., Van Maaren, M., Moncada-Torres, A., Geleijnse, G., & Jansen, T. (2020). Machine Learning Explainability in Breast Cancer Survival. Studies in Health Technology and Informatics, 270, 307–311. https://doi.org/10.3233/shti200172

Holzinger, A., Langs, G., Müller, H., Zatloukal, K., & Denk, H. (2019). Causability and explainability of artificial intelligence in medicine. WIREs Data Mining and Knowledge Discovery, 9(4). https://doi.org/10.1002/widm.1312

Hosaïn, M. T., Tabassum, R., Rafï, S., Siddiky, M. M., Insïa, K., & Anïk, M. H. (2023). Path To Gain Functional Transparency In Artificial Intelligence With Meaningful Explainability. Journal of Metaverse, 3(2), 166–180. https://doi.org/10.57019/jmv.1306685

Hudon, A., Léger, P.-M., Sénécal, S., Karran, A., & Demazure, T. (2021). Explainable Artificial Intelligence (XAI): How the Visualization of AI Predictions Affects User Cognitive Load and Confidence (pp. 237–246). Springer. https://doi.org/10.1007/978-3-030-88900-5_27

Islam, M. U., Alam, Z. I., Zobaed, S. M., Hassan, M., Fazle Rabby, M., & Mozaharul Mottalib, M. (2022). The Past, Present, and Prospective Future of XAI: A Comprehensive Review (pp. 1–29). Springer. https://doi.org/10.1007/978-3-030-96630-0_1

Karim, M. R., Lange, C., Islam, T., Decker, S., Shajalal, M., Rebholz-Schuhmann, D., Cochez, M., & Beyan, O. (2023). Explainable AI for Bioinformatics: Methods, Tools and Applications. Briefings in Bioinformatics, 24(5). https://doi.org/10.1093/bib/bbad236

Kaur, D., Rittichier, K. J., Uslu, S., & Durresi, A. (2022). Trustworthy Artificial Intelligence: A Review. ACM Computing Surveys, 55(2), 1–38. https://doi.org/10.1145/3491209

Khan, N., Nauman, M., Akhtar, N., Almadhor, A. S., Alghuried, A., & Alhudhaif, A. (2024). Guaranteeing Correctness in Black-Box Machine Learning: A Fusion of Explainable AI and Formal Methods for Healthcare Decision-Making. IEEE Access, 12, 90299–90316. https://doi.org/10.1109/access.2024.3420415

Kim, J., Maathuis, H., & Sent, D. (2024). Human-centered evaluation of explainable AI applications: a systematic review. Frontiers in Artificial Intelligence, 7. https://doi.org/10.3389/frai.2024.1456486

Korade, D. (2024). Unlocking Machine Learning Model Decisions: A Comparative Analysis of LIME and SHAP for Enhanced Interpretability. Journal of Electrical Systems, 20(2s), 598–613. https://doi.org/10.52783/jes.1480

Liu, H., Liu, X., Tang, J., Fan, W., Jain, A., Liu, Y., Wang, Y., Jain, S., & Li, Y. (2022). Trustworthy AI: A Computational Perspective. ACM Transactions on Intelligent Systems and Technology, 14(1), 1–59. https://doi.org/10.1145/3546872

Ma, X., Hou, M., Zhan, J., & Liu, Z. (2023). Interpretable Predictive Modeling of Tight Gas Well Productivity with SHAP and LIME Techniques. Energies, 16(9), 3653. https://doi.org/10.3390/en16093653

Madhav, A. V. S., & Tyagi, A. K. (2022). Explainable Artificial Intelligence (XAI): Connecting Artificial Decision-Making and Human Trust in Autonomous Vehicles (pp. 123–136). Springer Nature Singapore. https://doi.org/10.1007/978-981-19-1142-2_10

Marey, A., Arjmand, P., Alerab, A. D. S., Eslami, M. J., Saad, A. M., Sanchez, N., & Umair, M. (2024). Explainability, transparency and black box challenges of AI in radiology: impact on patient care in cardiovascular radiology. Egyptian Journal of Radiology and Nuclear Medicine, 55(1). https://doi.org/10.1186/s43055-024-01356-2

Mathew, D. E., Ebem, D. U., Ikegwu, A. C., Ukeoma, P. E., & Dibiaezue, N. F. (2025). Recent Emerging Techniques in Explainable Artificial Intelligence to Enhance the Interpretable and Understanding of AI Models for Human. Neural Processing Letters, 57(1). https://doi.org/10.1007/s11063-025-11732-2

Metta, C., Rinzivillo, S., Beretta, A., Giannotti, F., & Pellungrini, R. (2024). Towards Transparent Healthcare: Advancing Local Explanation Methods in Explainable Artificial Intelligence. Bioengineering, 11(4), 369. https://doi.org/10.3390/bioengineering11040369

Morato De Andrade, O., & Sousa Alves, M. A. (2024). Using Explainable Artificial Intelligence (XAI) to reduce opacity and address bias in algorithmic models. Revista Thesis Juris, 13(1), 03–25. https://doi.org/10.5585/13.2024.26510

Olateju, O. O., Samuel-Okon, A. D., Okon, S. U., Asonze, C. U., & Olaniyi, O. O. (2024). Exploring the Concept of Explainable AI and Developing Information Governance Standards for Enhancing Trust and Transparency in Handling Customer Data. Journal of Engineering Research and Reports, 26(7), 244–268. https://doi.org/10.9734/jerr/2024/v26i71206

Owolabi, O. S., Uche, P. C., Islam, R. B., Ihejirika, C., Adeniken, N. T., & Chhetri, B. J. T. (2024). Ethical Implication of Artificial Intelligence (AI) Adoption in Financial Decision Making. Computer and Information Science, 17(1), 49. https://doi.org/10.5539/cis.v17n1p49

Panda, M., & Mahanta, S. R. (2024). Explainable Artificial Intelligence for Healthcare Applications Using Random Forest Classifier with LIME and SHAP (pp. 89–105). Crc. https://doi.org/10.1201/9781003442509-6

Pavlidis, G. (2024). Unlocking the black box: analysing the EU artificial intelligence act's framework for explainability in AI. Law, Innovation and Technology, 16(1), 293–308. https://doi.org/10.1080/17579961.2024.2313795

Pawar, U., O'Reilly, R., O'Shea, D., & Rea, S. (2020, June 1). Explainable AI in Healthcare. https://doi.org/10.1109/cybersa49311.2020.9139655

Pham, T. (2025). Ethical and legal considerations in healthcare AI: innovation and policy for safe and fair use. Royal Society Open Science, 12(5). https://doi.org/10.1098/rsos.241873

R, J. (2024). Transparency in AI Decision Making: A Survey of Explainable AI Methods and Applications. Advances in Robotic Technology, 2(1), 1–10. https://doi.org/10.23880/art-16000110

Raz, A., Avnoon, N., Inbar, Y., Eyal, G., & Heinrichs, B. (2024). Prediction and explainability in AI: Striking a new balance? Big Data &amp; Society, 11(1). https://doi.org/10.1177/20539517241235871

Ridley, M. (2024). <scp>Human-centered</scp> explainable artificial intelligence: An Annual Review of Information Science and Technology (ARIST) paper. Journal of the Association for Information Science and Technology, 76(1), 98–120. https://doi.org/10.1002/asi.24889

Sabharwal, R., Cook, P., Wamba, S. F., & Miah, S. J. (2024). Extending application of explainable artificial intelligence for managers in financial organizations. Annals of Operations Research. https://doi.org/10.1007/s10479-024-05825-9

Salih, A. M., Radeva, P., Raisi-Estabragh, Z., Menegaz, G., Lekadir, K., Petersen, S. E., & Galazzo, I. B. (2024). A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME. Advanced Intelligent Systems, 7(1). https://doi.org/10.1002/aisy.202400304

Shahrzad, H., Miikkulainen, R., & Hodjat, B. (2025). EVOTER: Evolution of Transparent Explainable Rule-sets. ACM Transactions on Evolutionary Learning and Optimization, 5(2), 1–30. https://doi.org/10.1145/3702651

Thalpage, N. (2023). Unlocking the Black Box: Explainable Artificial Intelligence (XAI) for Trust and Transparency in AI Systems. Journal of Digital Art &amp; Humanities, 4(1), 31–36. https://doi.org/10.33847/2712-8148.4.1_4

Udegbe, F., Ebulue, C., Ekesiobi, C., & Ebulue, O. (2024). THE ROLE OF ARTIFICIAL INTELLIGENCE IN HEALTHCARE: A SYSTEMATIC REVIEW OF APPLICATIONS AND CHALLENGES. International Medical Science Research Journal, 4(4), 500–508. https://doi.org/10.51594/imsrj.v4i4.1052

Upadhyay, U., Gradisek, A., Iqbal, U., Dhar, E., Li, Y.-C., & Syed-Abdul, S. (2023). Call for the responsible artificial intelligence in the healthcare. BMJ Health & Care Informatics, 30(1), e100920. https://doi.org/10.1136/bmjhci-2023-100920

Veitch, E., & Alsos, O. A. (2021). Human-Centered Explainable Artificial Intelligence for Marine Autonomous Surface Vehicles. Journal of Marine Science and Engineering, 9(11), 1227. https://doi.org/10.3390/jmse9111227

Vimbi, V., Shaffi, N., & Mahmud, M. (2024). Interpreting artificial intelligence models: a systematic review on the application of LIME and SHAP in Alzheimer's disease detection. Brain Informatics, 11(1). https://doi.org/10.1186/s40708-024-00222-1

Wang, Y. (2024). A Comparative Analysis of Model Agnostic Techniques for Explainable Artificial Intelligence. Research Reports on Computer Science.

https://doi.org/10.37256/rrcs.3220244750

Weber, P., Carl, K. V., & Hinz, O. (2023). Applications of Explainable Artificial Intelligence in Finance—a systematic review of Finance, Information Systems, and Computer Science literature. Management Review Quarterly, 74(2), 867–907. https://doi.org/10.1007/s11301-023-00320-0

Wolf, C. T., & Ringland, K. E. (2020). Designing accessible, explainable AI (XAI) experiences. ACM SIGACCESS Accessibility and Computing, 125(125), 1. https://doi.org/10.1145/3386296.3386302

Xu, Q., Xie, W., Liao, B., Hu, C., Qin, L., Yang, Z., Xiong, H., Lyu, Y., Zhou, Y., & Luo, A. (2023). Interpretability of Clinical Decision Support Systems Based on Artificial Intelligence from Technological and Medical Perspective: A Systematic Review. Journal of Healthcare Engineering, 2023(1), 1–13. https://doi.org/10.1155/2023/9919269

Zahoor, K., Zakaria Bawany, N., & Qamar, T. (2024). Evaluating text classification with explainable artificial intelligence. IAES International Journal of Artificial Intelligence (IJ-AI), 13(1), 278. https://doi.org/10.11591/ijai.v13.i1.pp278-286

Zhang, Y., Weng, Y., & Lund, J. (2022). Applications of Explainable Artificial Intelligence in Diagnosis and Surgery. Diagnostics, 12(2), 237. https://doi.org/10.3390/diagnostics12020237

Zhu, J., Risi, S., Liapis, A., Youngblood, G. M., & Bidarra, R. (2018). Explainable AI for Designers: A Human-Centered Perspective on Mixed-Initiative Co-Creation. 1–8. https://doi.org/10.1109/cig.2018.8490433