World Journal of Advanced Engineering Technology and Sciences
eISSN: 2582-8266
Cross Ref DOI: 10.30574/wjaets
Journal homepage: https://wjaets.com/

(REVIEW ARTICLE)

Check for updates

Explainable AI (XAI) for trustworthy and transparent decision-making: A theoretical framework for AI interpretability

Arunraju Chinnaraju *

Doctorate in Business Administration Student, Westcliff University, College of Business, California, USA.

World Journal of Advanced Engineering Technology and Sciences, 2025, 14(03), 170-207

Publication history: Received on 19 January 2025; revised on 03 March 2025; accepted on 05 March 2025

Article DOI: https://doi.org/10.30574/wjaets.2025.14.3.0106

Abstract

Explainable Artificial Intelligence (XAI) has become a critical area of research in addressing the black-box nature of complex AI models, particularly as these systems increasingly influence high-stakes domains such as healthcare, finance, and autonomous systems. This study presents a theoretical framework for AI interpretability, offering a structured approach to understanding, implementing, and evaluating explainability in AI-driven decision-making. By analyzing key XAI techniques, including LIME, SHAP, and DeepLIFT, the research categorizes explanation methods based on scope, timing, and dependency on model architecture, providing a novel taxonomy for understanding their applicability across different use cases. Integrating insights from cognitive theories, the framework highlights how human comprehension of AI decisions can be enhanced to foster trust and reliability. A systematic evaluation of existing methodologies establishes critical explanation quality metrics, considering factors such as fidelity, completeness, and user satisfaction. The findings reveal key trade-offs between model performance and interpretability, emphasizing the challenges of balancing accuracy with transparency in real-world applications. Additionally, the study explores the ethical and regulatory implications of XAI, proposing standardized protocols for ensuring fairness, accountability, and compliance in AI deployment. By providing a unified theoretical framework and practical recommendations, this research contributes to the advancement of explainability in AI, paving the way for more transparent, interpretable, and human-centric AI systems.

Keywords: Explainable Artificial Intelligence (XAI); Model Interpretability; Decision Transparency; Machine Learning; AI Ethics; Human-Ai Interaction; AI Accountability & Trustworthy

1. Introduction

Explainable Artificial Intelligence (XAI) has emerged as a transformational paradigm in AI development, addressing the growing demand for transparency, interpretability, and accountability in modern machine learning (ML) and deep learning (DL) models. As AI-driven systems increasingly make critical decisions in high-stakes domains such as healthcare, finance, law, and autonomous systems, concerns regarding their black-box nature have intensified (Lee et al., 2022). Traditional AI models, particularly deep learning architectures such as Convolutional Neural Networks (CNNs), Transformer-based models, and Reinforcement Learning systems, often operate in a manner that lacks human interpretability. The opacity of these models raises fundamental concerns about bias, fairness, ethical compliance, and decision accountability, especially when AI decisions influence regulatory policies, financial transactions, and human lives (Sun et al., 2024).

The Need for Explainability in AI Systems: The rise of AI-powered decision-making has fundamentally shifted the way organizations and individuals interact with technology. However, the complexity of modern AI models has made it difficult to understand, justify, and audit AI-generated decisions. This lack of explainability poses several risks: Lack of

* Corresponding author: Arunraju Chinnaraju.

User Trust: If users cannot understand how an AI system reaches its conclusions, they may be hesitant to rely on its recommendations, especially in sensitive applications such as medical diagnostics or financial risk assessment (Neves et al., 2023). Regulatory and Legal Compliance: Many governments and regulatory bodies, including the European Union's General Data Protection Regulation (GDPR) and the U.S. Federal Trade Commission (FTC), now require AI systems to provide human-interpretable justifications for automated decisions. Ethical Considerations and Bias Mitigation: Without proper interpretability, AI models may perpetuate biases present in the training data, leading to unfair outcomes in areas such as hiring, lending, and criminal justice (Knof et al., 2024).

To address these challenges, Explainable AI (XAI) provides a structured approach to making AI decision-making transparent, interpretable, and accountable. XAI techniques can be broadly categorized into intrinsic interpretability and post-hoc interpretability, each playing a unique role in AI system design.

1.1. Intrinsic vs. post-Hoc Explainability

Intrinsic Interpretability: Designing AI Systems for Transparency. Intrinsic interpretability refers to AI models that are inherently explainable due to their architecture and simplicity. These models are designed to be transparent by default, making them easier to understand and audit (Vale et al., 2022). Examples include: Decision Trees: A tree-based structure that clearly outlines how AI reaches decisions by splitting data into interpretable nodes. Generalized Additive Models (GAMs): A type of regression-based model that provides clear feature importance insights, ensuring decision logic is understandable. Rule-Based Systems: These models follow predefined rules, often used in expert systems and legal AI applications (Mota et al., 2024). Although intrinsically interpretable models offer high transparency, they often lack the complexity needed for highly nonlinear, high-dimensional problems such as image recognition or natural language processing (NLP).

Post-Hoc Explainability: Interpreting Black-Box Models. In contrast, post-hoc explainability methods aim to explain already-trained AI models without altering their internal mechanisms. These techniques provide insight into how a complex AI model makes decisions while maintaining the model's predictive power (Retzlaff et al., 2024).

Key post-hoc explainability methods include: Local Interpretable Model-Agnostic Explanations (LIME): This method generates localized surrogate models that approximate the decision boundary of a black-box AI model. By perturbing input data and observing changes in predictions, LIME creates an interpretable linear model that explains individual AI decisions. SHapley Additive exPlanations (SHAP): Based on cooperative game theory, SHAP assigns contribution values to each input feature, ensuring a fair and consistent explanation of how different factors influenced an AI's prediction (Narkhede, 2024). Layer-wise Relevance Propagation (LRP): A deep-learning-specific technique that traces back AI decisions layer by layer, identifying which features or neurons contributed most to an outcome. LRP is widely used in computer vision applications. Post-hoc interpretability has gained significant traction due to its scalability and applicability to deep learning models. However, challenges remain in ensuring that explanations are faithful to the actual reasoning process of the AI model and do not introduce misleading approximations (Belaid et al., 2023).

1.2. Evolution from Black-Box Models to Transparent AI Systems

The development of black-box AI models, particularly deep learning architectures, has significantly advanced predictive accuracy across industries. However, early deep learning models, including Multi-Layer Perceptron's (MLPs), Recurrent Neural Networks (RNNs), and Deep Neural Networks (DNNs), were highly opaque. The shift toward explainability has led to the emergence of more interpretable deep-learning architectures, including: Attention Mechanisms (e.g., Transformer-based Models like BERT & GPT): These models highlight the most relevant parts of an input sequence, making their decision-making process partially explainable. Graph Neural Networks (GNNs): Used in domains such as fraud detection and bioinformatics, GNNs incorporate relational data into their decision-making process, improving interpretability. Hybrid AI Systems: Combining symbolic AI with deep learning to balance accuracy and explainability (Bhagavatula et al., 2024).

Applications of XAI in High-Stakes Domains: Explainability is particularly important in high-risk AI applications, where decisions must be justifiable, auditable, and ethically sound. Some key areas include: Healthcare: AI-powered medical diagnosis tools must explain why a certain disease is detected in an X-ray scan to help doctors validate the results (Tian et al., 2023). Techniques such as Grad-CAM (Gradient-weighted Class Activation Mapping) visualize which regions of an image influenced a model's decision. Finance: Interpretable AI models in financial risk assessment are essential for ensuring compliance with regulations such as the Fair Credit Reporting Act (FCRA) and avoiding discrimination in lending decisions (Friesel & Spinczyk, 2022). Legal Systems: AI-driven legal analytics tools are used for case law analysis and sentencing predictions, requiring transparent decision rationales to prevent biases.

Challenges in Achieving AI Transparency: Despite its advancements, achieving AI transparency remains a complex challenge. The main obstacles include: Trade-off Between Accuracy and Interpretability: Deep learning models often achieve superior accuracy but are difficult to explain. Simplifying them for transparency may compromise predictive power. Scalability of Explainability Methods: Many post-hoc techniques, such as SHAP and LIME, become computationally expensive when applied to large-scale datasets (Akhai, 2023). Human-Centric Explanations: AI explanations must align with human cognitive models, ensuring that users—whether experts or non-experts—can understand and trust the reasoning behind AI decisions.

Standardization and Regulatory Compliance in XAI: As AI systems become more integrated into society, governance, and business, regulatory bodies have started to demand explainability as a fundamental requirement. The European Commission's AI Act, NIST's Explainable AI Guidelines, and IEEE's AI Ethics Standards all emphasize the need for standardized XAI evaluation metrics (Tian et al., 2023). Standardization efforts focus on: Defining Benchmarks for Explanation Quality: Ensuring that explanations are accurate, complete, and reliable. Ensuring Cross-Domain Consistency: Creating generalizable frameworks that work across healthcare, finance, and other industries. Improving User Interpretability: Designing AI interfaces that make explanations intuitive and actionable (Bhagavatula et al., 2024).

The increasing emphasis on explainability in AI systems has led to rapid advancements in XAI methodologies. The following section explores the historical evolution of XAI, tracing its development from early rule-based systems to modern deep-learning explanation techniques. Additionally, it highlights existing research gaps and emerging trends that shape the future of explainable AI.

2. Literature Review

Artificial Intelligence (AI) has become a dominant force in decision-making processes across numerous domains, from healthcare and finance to autonomous systems and legal applications. While deep learning models have revolutionized predictive accuracy, they are inherently opaque, often functioning as black-box systems that provide little to no insight into their internal decision-making processes. The rise of Explainable Artificial Intelligence (XAI) seeks to address this issue by developing methods that make AI more interpretable while maintaining high performance (Bhagavatula et al., 2024). This literature review provides an in-depth theoretical analysis of the evolution of XAI, its key techniques, challenges in implementation, and existing research gaps, offering a critical assessment of previous implementations of interpretability methods

2.1. Historical Evolution of XAI

The demand for AI explainability has been evident since the early days of AI research. The evolution of XAI can be understood by examining its progression from rule-based expert systems to modern deep learning interpretability techniques.

Early Rule-Based Expert Systems (1970s–1990s): The earliest AI systems were built using rule-based expert systems, where human knowledge was encoded into structured if-then rules. Notable examples include MYCIN (for medical diagnosis), DENDRAL (for chemical analysis), and PROSPECTOR (for mineral exploration). These systems were highly interpretable because they explicitly outlined their reasoning process, allowing users to trace the decision path (Kozielski et al., 2025). However, their reliance on manually crafted rules limited their scalability and adaptability. As problems became more complex and required large-scale data processing, rule-based systems struggled to provide generalizable solutions. This led to the emergence of statistical machine learning models in the 1990s (Jang et al., 2023).

Emergence of Statistical Machine Learning (1990s–2000s): The shift from symbolic AI to data-driven learning models marked a major transition in AI research. Algorithms such as Support Vector Machines (SVMs), Decision Trees, and Random Forests became popular due to their ability to learn patterns directly from data (Ramachandran, 2023). Some of these models, like decision trees, offered inherent interpretability, but others, such as SVMs, were significantly less transparent. During this period, AI systems became more autonomous, but explainability was no longer a priority, as emphasis shifted toward improving accuracy and efficiency (Shojaeinasab et al., 2024). This trade-off set the stage for the black-box problem that emerged with deep learning.

Rise of Deep Learning and the Black-Box Problem (2010s–Present): The explosion of deep learning in the 2010s, particularly with the development of Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformer-based architectures, led to state-of-the-art performance in multiple AI applications (Dağlarli, 2020). These models were capable of detecting patterns in large datasets with unprecedented accuracy, but they also introduced

extreme opacity. Unlike rule-based or traditional ML models, deep learning networks consisted of millions of parameters, making it virtually impossible for humans to interpret their decisions (Kosov et al., 2024).

Concerns about AI transparency became a regulatory issue following the introduction of the General Data Protection Regulation (GDPR) in 2018, which established a "right to explanation" for automated decision-making. This led to a renewed focus on post-hoc interpretability techniques and efforts to develop more transparent AI architectures (YazdaniBanafsheDaragh & Malek, 2021). Today, XAI research is centered around two key areas: intrinsic interpretability (building inherently explainable models) and post-hoc explainability (developing techniques to explain black-box models after training).

2.2. Comparative Analysis of XAI Techniques

XAI methods can be categorized into two main approaches: intrinsically interpretable models and post-hoc explainability techniques. Some AI models are designed to be inherently interpretable, meaning their decision-making logic is transparent by default. These models often trade off predictive power for increased explainability.

2.2.1. Intrinsically Interpretable Models

Generalized Additive Models (GAMs): Generalized Additive Models (GAMs) extend linear regression by allowing non-linear transformations of features while preserving a high level of interpretability. GAMs are particularly useful in applications where transparency is critical, such as healthcare and finance. In healthcare, for example, GAM-based models have been successfully used to predict disease risk factors while providing doctors with an explicit understanding of how each feature contributes to the diagnosis (Mokdad et al., 2024). However, GAMs struggle with high-dimensional data and cannot effectively capture complex feature interactions, making them unsuitable for deep learning tasks. Decision Trees and Rule-Based Systems: Decision Trees, including Random Forests and XGBoost, provide a hierarchical structure where each decision is based on a sequence of feature splits. These models are widely used in fraud detection, risk assessment, and medical diagnostics due to their transparency. However, as decision trees grow deeper, they become harder to interpret, leading to the "complexity vs. interpretability trade-off." Additionally, rule-based systems, while interpretable, lack adaptability, making them impractical for dynamic real-world environments (B.R & V, 2024).

2.2.2. Post-Hoc Explainability Techniques

Post-hoc explainability techniques seek to provide interpretations for already trained black-box models without altering their internal mechanisms. These methods help make AI systems more transparent to stakeholders, but each technique has limitations that affect reliability, scalability, and real-world applicability.

Local Interpretable Model-Agnostic Explanations (LIME): LIME generates explanations by perturbing input data and observing how predictions change, fitting a local interpretable model (e.g., linear regression) to approximate the black-box decision boundary. While LIME is widely used in finance, healthcare, and law enforcement, it suffers from stability issues that impact reliability (Kaushik et al., 2024). One major limitation of LIME is that it does not guarantee consistency across similar inputs. Because the method relies on random perturbations, different runs may produce different explanations for the same prediction, making it difficult to establish trust in high-stakes applications. For example, in medical AI, an AI-driven cancer diagnosis model explained by LIME may present different feature importances for the same patient scan across multiple runs, leading to uncertainty and lack of confidence among physicians. This inconsistency undermines LIME's usability in regulated industries that require auditability and accountability in AI decision-making. Additionally, LIME assumes that a linear model can approximate complex non-linear decision boundaries, which is often unrealistic for highly non-linear models such as deep learning architectures (Kamath & Liu, 2021). This oversimplification means that LIME's explanations may misrepresent how the actual model makes decisions, leading to potentially misleading interpretations.

SHapley Additive exPlanations (SHAP): SHAP is considered one of the most theoretically sound explainability methods, leveraging game-theoretic principles to fairly distribute feature importance scores. Unlike LIME, SHAP ensures consistency and additivity, making it a more reliable technique for high-risk applications. However, SHAP's major drawback is its computational inefficiency, which limits its practicality in real-time and large-scale AI systems (Vale et al., 2022). A key issue with SHAP is that it requires computing feature attributions across all possible feature subsets, a process that becomes exponentially complex as the number of features increases. In deep learning applications, where models often involve hundreds or thousands of input features, SHAP can take an unreasonable amount of time to generate explanations (Mota et al., 2024). For example, in fraud detection, a model evaluating thousands of transaction variables may take hours or even days to compute SHAP values for a single decision, making it infeasible for real-time

analysis. Another challenge is that while SHAP provides both local and global explanations, its global explanations tend to be less intuitive for end-users. Business stakeholders, doctors, or legal professionals may find it difficult to interpret a feature's global importance across an entire dataset, whereas localized case-by-case explanations may be more useful. This highlights the ongoing tension between technical accuracy and practical usability in XAI.

Layer-wise Relevance Propagation (LRP): LRP is designed specifically for deep neural networks and works by propagating prediction relevance scores backward through the network layers to determine which features contributed most to the final decision. LRP has been particularly successful in computer vision applications, such as explaining CNN-based medical image classification models. Despite its success, LRP has several limitations. First, it is highly dependent on the model's architecture, meaning its effectiveness varies significantly between different network structures (Miró-Nicolau et al., 2024). Unlike SHAP and LIME, which can be applied to various ML models, LRP is restricted to deep learning models and requires custom configurations for each architecture. Second, LRP can be unstable in high-dimensional input spaces, where small perturbations in input data can lead to drastic changes in explanation outputs. This is a significant concern in applications like autonomous driving, where slight variations in an image (e.g., lighting changes) might lead to different heatmaps in LRP explanations, raising concerns about robustness and reliability. These challenges suggest that while post-hoc methods provide valuable insights, they do not fully solve the AI interpretability problem and require improvements in stability, efficiency, and model-agnostic applicability (Narkhede, 2024).

2.3. Challenges in XAI Implementation

Scalability and Real-Time Feasibility: Many XAI techniques, particularly SHAP and LIME, require intensive computational resources, making them impractical for real-time AI systems. Industries like automated trading, fraud detection, and autonomous systems demand split-second decision-making, and existing XAI methods often fail to meet such speed requirements (Kaushik et al., 2024). Scaling explainability to handle deep learning models in real-time remains a fundamental challenge.

The Accuracy-Interpretability Trade-Off: A fundamental issue in XAI research is the trade-off between model accuracy and interpretability. While deep learning models provide unparalleled accuracy, they are inherently less explainable (Vale et al., 2022). Conversely, inherently interpretable models (such as GAMs and decision trees) offer transparency at the cost of predictive performance. Striking a balance between these two properties remains an ongoing challenge.

Lack of Standardized Evaluation Metrics: Currently, there is no universally accepted framework for assessing XAI effectiveness. Different studies measure explainability using inconsistent criteria, making it difficult to compare techniques (Mota et al., 2024). Developing a standardized benchmarking system for evaluating XAI methods is crucial for ensuring reliability and trust.

User-Specific Explainability Needs: AI explanations must be tailored to different audiences. Regulatory bodies, domain experts, and lay users require different levels of explanation detail. Current XAI techniques do not adequately adapt explanations to different stakeholder needs, limiting their practical adoption in industry (Dağlarli, 2020).

Lack of Causal Explanations: Most XAI techniques focus on correlation-based explanations rather than causal reasoning. Understanding why a model made a specific decision, rather than just how input features contributed, is crucial for trust and accountability. There is growing interest in integrating causal inference models into XAI research (Jang et al., 2023).

2.4. Research Gaps and Future Directions

Developing Scalable, High-Fidelity Explanation Techniques: Current XAI methods, particularly post-hoc techniques, are computationally expensive and slow (Kosov et al., 2024). Future research should focus on scalable, high-fidelity explainability approaches that can work in real-time AI applications without compromising accuracy.

Integration of Causal Inference for Meaningful Explanations: XAI methods must move beyond correlation-based feature attributions to incorporate causal inference (Ramachandran, 2023). Causal models would allow AI systems to justify decisions in a way that aligns with human reasoning, improving accountability.

Standardizing XAI Evaluation Metrics: The lack of a common evaluation framework hinders XAI research progress. Future work must develop standardized benchmarks and regulatory guidelines to assess explanation quality across different domains and models (Kozielski et al., 2025).

Developing User-Centric Explainability Models: Most XAI techniques assume a one-size-fits-all approach to explainability, but real-world applications require tailored explanations based on user expertise levels (Akhai, 2023).

Research should focus on developing adaptive AI models that personalize explanations based on the end user's domain knowledge.

Given these challenges and research gaps, the next section introduces a theoretical framework for integrating explainability into AI systems. This framework will define key principles, evaluation criteria, and methodologies necessary to ensure AI is not only accurate but also transparent, interpretable, and trustworthy.

3. Theoretical Framework for Explainable AI (XAI)

The development of a theoretical framework for Explainable Artificial Intelligence (XAI) is fundamental to ensuring that AI-driven systems operate in a transparent, interpretable, and accountable manner. As AI technologies become deeply integrated into decision-making processes across industries such as healthcare, finance, and law, their black-box nature raises critical concerns regarding fairness, bias, ethical considerations, and regulatory compliance (Açar, 2022). Theoretical foundations in XAI provide structured methodologies that not only define the principles of explainability but also establish guidelines for designing, evaluating, and deploying AI models that align with human cognitive reasoning and decision-making expectations.

The core aspect of XAI revolves around the distinction between local and global explanations, which serve different interpretability needs. Local explanations focus on individual AI decisions, allowing users to understand the specific factors influencing a particular outcome, while global explanations aim to provide an overarching view of how an AI system derives patterns and decisions across all inputs (Arun Sampaul Thomas et al., 2024). Local explanations are particularly valuable in contexts where stakeholders require transparency for case-by-case decision-making, such as loan approvals, medical diagnoses, and fraud detection, whereas global explanations help AI developers, regulators, and domain experts evaluate biases, consistency, and fairness in model behavior across datasets. A well-defined XAI framework must incorporate both forms of explanations to ensure AI systems remain comprehensible and auditable.

Another critical consideration in the theoretical framework is the distinction between intrinsic and post-hoc explainability. Intrinsic explainability refers to models that are inherently interpretable by design, such as decision trees, linear regression models, and Generalized Additive Models (GAMs). These models enable direct reasoning from input variables to predictions, making them preferable for applications requiring immediate justification (Chauhan et al., 2024). Post-hoc explainability, on the other hand, is necessary for complex black-box models such as deep neural networks, transformer architectures, and ensemble learning models, which demand interpretability techniques applied after training to rationalize their decisions. Post-hoc explainability methods include SHAP, LIME, Grad-CAM, and Layer-wise Relevance Propagation (LRP), which approximate how the AI model derives its predictions. The challenge within XAI is ensuring that post-hoc explanations faithfully represent the actual decision process rather than merely providing approximations that may not align with the true underlying mechanics of the model (Gummadi et al., 2024).

The model-specific versus model-agnostic distinction further refines the theoretical approach to explainability. Model-specific methods are tailored to specific AI architectures, leveraging the internal computational properties of the model to generate explanations (Akhai, 2024). For example, Grad-CAM is an explainability method specifically designed for Convolutional Neural Networks (CNNs), visualizing the regions of an image that influenced a classification decision. Layer-wise Relevance Propagation (LRP) operates similarly for deep neural networks by tracing back neuron activations to identify the most influential input features. In contrast, model-agnostic methods such as SHAP and LIME provide universal interpretability techniques that can be applied to any AI model, regardless of its underlying architecture. While model-agnostic methods enhance flexibility, they often lack the precision of model-specific techniques, making them less effective in explaining highly complex models with deep hierarchical structures (Explainable AI (XAI): Bridging the Gap between Machine Learning and Human Understanding, 2024).

A well-established XAI framework must also integrate cognitive theories and human-centered explanation approaches to enhance the usability and adoption of AI explanations. Human cognition relies on causal reasoning and mental model formation, meaning that AI explanations must align with how people naturally process information. If an AI system provides an overly technical or non-intuitive explanation, end-users may fail to trust or understand the decision-making process, even if the explanation is mathematically sound. Research in cognitive psychology and human-computer interaction (HCI) suggests that counterfactual reasoning and contrastive explanations which describe how a decision would have changed under different input conditions are more effective for human comprehension. A theoretical XAI framework should therefore emphasize the importance of generating explanations that align with human cognitive patterns and decision-making behaviors (Palacio et al., 2021).

Incorporating causality into AI explanations is another significant aspect of the theoretical foundation of XAI. Most existing interpretability techniques rely on correlation-based feature attributions, where models highlight which variables contributed the most to an outcome. However, this does not necessarily imply causation, meaning that the true underlying factors driving AI decisions may remain hidden. By integrating causal inference models, AI explanations can shift from descriptive correlations to actionable causal insights, providing more meaningful and trustworthy justifications for AI-driven outcomes (Hughes et al., 2020). This shift is particularly relevant in healthcare, legal systems, and autonomous systems, where AI decisions must align with causal reasoning frameworks to ensure fairness and prevent spurious correlations from influencing high-stakes decision-making.

Another important component of the theoretical framework is the role of counterfactual explanations in enhancing interpretability. Counterfactual reasoning is essential in understanding how an AI decision would change if certain input features were altered (D et al., 2024). For instance, in a loan approval model, a counterfactual explanation might state: "If the applicant had a 10% higher income, they would have qualified for the loan." Such explanations provide actionable insights that allow users to interpret AI decisions in a more intuitive and relatable manner. They also serve regulatory and fairness objectives, enabling AI developers and policymakers to assess whether an AI system systematically disadvantages certain demographic groups by analyzing how decision thresholds shift across populations.

The theoretical foundation of XAI also necessitates a clear distinction between transparency, interpretability, and explainability, as these terms are often used interchangeably despite their nuanced differences. Transparency refers to the extent to which an AI system's architecture, training data, and decision-making processes are open and accessible (Tchuente et al., 2024). Interpretability focuses on the degree to which a human can understand the relationship between input features and model outputs. Explainability extends beyond interpretability by ensuring that AI decisions are not only understandable but also actionable and meaningful to stakeholders. A comprehensive XAI framework must balance these three dimensions to ensure that AI explanations are both technically robust and practically useful across different stakeholder groups (Kumari et al., 2022).

Another significant component of the XAI theoretical framework is the integration of regulatory and ethical principles into explainability methodologies. AI models deployed in finance, healthcare, and law enforcement are increasingly subjected to regulatory frameworks such as GDPR's "right to explanation" provision, which mandates that automated decision-making systems provide clear and justifiable reasoning for their outputs (Goenka et al., 2024). AI systems must also adhere to fairness constraints, ensuring that their decision-making processes do not reinforce societal biases. A strong theoretical XAI foundation must therefore incorporate fairness-aware interpretability techniques, such as bias detection algorithms and adversarial de-biasing methods, which analyze whether AI models are making equitable decisions across different demographic groups (B.R & V, 2024).

Finally, the theoretical framework must outline evaluation metrics and validation methodologies for assessing explainability in AI systems. While several explainability techniques exist, there is no universally accepted standard for measuring the effectiveness of AI explanations. Researchers have proposed multiple criteria, such as fidelity (how accurately an explanation represents model behavior), completeness (how well an explanation accounts for all influencing factors), and consistency (how stable explanations remain across similar predictions). However, the absence of an industry-wide benchmarking system means that evaluating and comparing XAI techniques remains subjective (Sharma et al., 2020). A robust XAI framework should establish standardized evaluation protocols to ensure that explainability methods are rigorously tested across multiple AI architectures and application domains.

3.1. Structured Process for XAI: The Five-Stage Theoretical Framework

The Five-Stage Theoretical Framework for XAI establishes a structured, repeatable process that ensures explainability is not an afterthought but an integrated component of AI development. This framework consists of five interconnected stages, ensuring that AI explanations are meaningful, scalable, and aligned with both technical and human-centered objectives.

3.1.1. Stage 1: Explanation Objective Definition

The first step in the XAI framework is to define the objective of explainability based on the needs of different stakeholders. AI users include model developers, regulators, domain experts, and general end-users, each requiring different types of explanations. Defining the objective clarifies whether explainability is required for technical validation, regulatory compliance, ethical considerations, or user trust-building (Chauhan et al., 2024). A well-defined explainability objective ensures that AI systems provide the right type of interpretation to the right audience. For example, in healthcare AI, a doctor using an AI-driven diagnosis tool requires local explanations to understand why a particular patient was diagnosed with a condition, whereas a hospital administrator might need global explanations to

assess whether the model is biased toward certain demographic groups. Similarly, in financial applications, a bank loan applicant may require a straightforward justification for why their loan was denied, while regulators need an audit trail to determine if the AI model systematically discriminates against certain demographics. This stage establishes the fundamental purpose of AI transparency, guiding the selection of appropriate interpretability techniques (Tchuente et al., 2024).

3.1.2. Selection of Explanation Type

Once the objective is defined, the framework requires selecting the appropriate type of explanation. The two primary categories are local explanations and global explanations. Local explanations focus on interpreting individual AI predictions to provide case-specific justifications. Techniques such as SHAP, LIME, and Counterfactual Explanations are commonly used to highlight the most influential factors behind a single decision (Gummadi et al., 2024). These explanations are particularly relevant in high-stakes applications such as medical diagnostics, legal AI systems, and fraud detection, where stakeholders must understand the specific reasons behind each prediction. Conversely, global explanations provide an overarching understanding of the model's general behavior across datasets. Methods such as Partial Dependence Plots (PDPs), Feature Importance Analysis, and Accumulated Local Effects (ALE) allow analysts to observe general patterns in how different variables influence predictions. While local explanations help individual users assess specific cases, global explanations are crucial for model debugging, ensuring fairness, and regulatory compliance. At this stage, the framework determines whether an explanation should be tailored to individual cases or should reflect broader model insights (B.R & V, 2024). In AI-driven financial systems, for example, an individual applicant's loan rejection requires a local explanation, whereas an institution assessing fairness across different racial groups requires a global explanation.

3.1.3. Selection of Explanation Methodology

The third stage involves selecting the appropriate explanation technique based on whether the AI model is intrinsically interpretable or requires post-hoc interpretability. Intrinsic interpretability refers to models that are inherently explainable, such as Decision Trees, Generalized Additive Models (GAMs), and Linear Regression, which enable direct reasoning from input variables to predictions. These models are preferred when transparency is a primary requirement (Palacio et al., 2021). However, for more complex black-box models, such as deep neural networks, transformer architectures, and ensemble learning models, explanations must be generated after training through post-hoc explainability techniques. Post-hoc interpretability techniques can be categorized into model-specific methods and model-agnostic methods. Model-specific methods leverage a model's internal structure to provide explanations, such as Grad-CAM for CNN-based image classification or Layer-wise Relevance Propagation (LRP) for deep learning networks. These techniques produce faithful interpretations by using model-inherent information (Kumari et al., 2022). Meanwhile, model-agnostic methods such as SHAP and LIME offer flexible interpretability solutions applicable to any AI model, regardless of its underlying architecture. While model-agnostic techniques enhance generalizability, they may fail to capture the true internal decision logic of a black-box model. At this stage, the framework ensures that the chosen explanation method aligns with the interpretability goals identified in Stage 1, balancing the trade-offs between explanation fidelity, computational efficiency, and user comprehensibility (Hughes et al., 2020).

3.1.4. Human-Centric Explanation Alignment

Technical explainability alone is insufficient unless it aligns with human cognitive processes. AI explanations must be structured in a way that is intuitive, useful, and actionable for different stakeholders. This stage of the framework ensures that explanations are adapted to user expertise levels and decision-making behaviors. The effectiveness of AI explanations is highly dependent on how well they align with human reasoning models, making it critical to incorporate principles from cognitive science and human-computer interaction (HCI). Research suggests that counterfactual reasoning and contrastive explanations which describe how a decision would have changed under different input conditions are more effective for human comprehension (Akhai, 2024). Moreover, explanations should minimize cognitive overload, ensuring that users receive just the right amount of information needed to make informed decisions without being overwhelmed. AI systems in healthcare, legal, and financial applications should generate interactive and tailored explanations, where domain experts receive detailed model insights, while non-experts receive simplified justifications (Goenka et al., 2024).

3.1.5. Evaluation and Validation of Explainability Methods

The final stage ensures that AI explanations meet quality, fairness, and regulatory standards. Since different interpretability methods produce different outputs, rigorous validation is necessary to assess whether explanations are consistent, reliable, and practically useful. The framework evaluates explainability techniques across three key dimensions: fidelity, user comprehension, and fairness detection. Fidelity assessment ensures that explanations

accurately reflect the true decision-making process of the AI model, rather than offering simplistic approximations (Chauhan et al., 2024). Metrics such as feature attribution consistency, completeness, and stability help determine whether an explanation method reliably captures model behavior. User comprehension testing evaluates whether explanations are understandable, actionable, and effective for different stakeholder groups. For example, AI-driven medical diagnosis models must ensure that doctors can correctly interpret explanations to validate predictions and improve patient outcomes. Fairness and bias detection mechanisms ensure that AI models do not reinforce discriminatory biases (Gummadi et al., 2024). Methods such as Demographic Parity Analysis and Adversarial De-biasing Checks are used to assess whether explanations remain equitable across different demographic groups. This stage of the framework establishes standardized evaluation methodologies to ensure trustworthy and regulatory-compliant AI explanations.

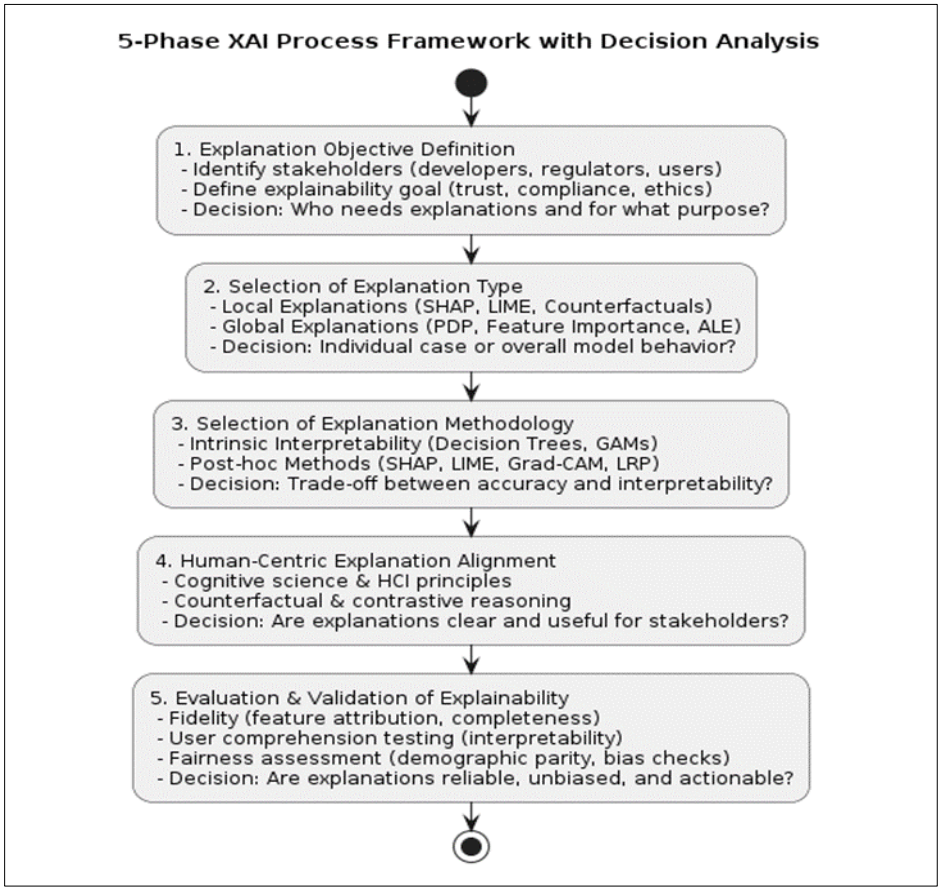

Figure 1 Five Phase XAI Implementation framework with Decision Analysis

The Five-Stage Theoretical Framework for XAI provides a structured approach to ensuring that AI systems are transparent, interpretable, and human-centric. By defining clear objectives, selecting appropriate explanation techniques, aligning explanations with cognitive processes, and rigorously validating interpretability methods, AI practitioners can systematically integrate explainability into AI development. However, explainability methods differ significantly in technical implementation, scalability, and effectiveness across different AI architectures (D et al., 2024). The next section will explore specific methods for interpreting AI decisions, detailing how various interpretability techniques are applied in real-world AI models and their practical implications across industries.

4. Methods for Interpreting AI Decisions

As AI systems increasingly influence decision-making across various high-stakes domains, the ability to interpret and explain their decisions has become a critical requirement. While machine learning models, particularly deep learning architectures, excel in predictive performance, their inherent complexity often renders them opaque and difficult to interpret (Mokdad et al., 2024). This opacity raises concerns regarding trust, accountability, fairness, and bias detection, necessitating the development of explainability techniques that can provide insight into model decision-making processes. Various interpretability methods have been developed to analyze and explain AI model behavior, each

offering unique advantages depending on the context of deployment and the level of interpretability required. These methods fall into distinct categories, including feature attribution techniques, perturbation-based approaches, visualization-based methods, rule-based interpretations, surrogate modeling, and counterfactual explanations (Kamath & Liu, 2021). The choice of an appropriate interpretability method depends on the model complexity, the type of data being analyzed, and the specific needs of the end-user.

Table 1 Factors influencing between Post-hoc Vs Intrinsically Interpretable Model

Factor	Post-Hoc Explainability	Intrinsically Interpretable Models
Performance	Works with high-accuracy models (deep learning, ensembles).	Limited to simple models (linear regression, decision trees).
Flexibility	Can be applied to any black-box model.	Model needs to be explicitly designed for interpretability.
Explainability Quality	Provides explanations, but may introduce approximations.	Directly interpretable, no need for external methods.
Use Cases	Widely used in real-world applications requiring accuracy + transparency.	Used where transparency is more important than accuracy.
Challenges	Can be computationally expensive and unstable (e.g., LIME's instability, SHAP's complexity).	Less powerful for complex tasks, struggles with high-dimensional data.

4.1. Intrinsically Interpretable Models

Intrinsically interpretable models are designed to provide transparency in their decision-making process by inherently offering explanations without the need for additional post-hoc explainability techniques. Unlike complex black-box models such as deep neural networks, which require external methods like SHAP or LIME to generate explanations, intrinsically interpretable models allow direct insight into the relationship between input features and predictions (Vale et al., 2022). These models are particularly valuable in high-stakes decision-making environments, such as healthcare, finance, and legal systems, where accountability, fairness, and trust are essential. By ensuring interpretability at the model level, they help regulatory bodies, domain experts, and end-users understand, validate, and audit AI-driven decisions effectively (Mota et al., 2024).

4.1.1. Linear Models: The Fundamental Basis of Interpretability

Linear models are among the most basic yet powerful intrinsically interpretable models, making them widely used in data science and machine learning. Linear regression, the simplest of these models, assumes a direct linear relationship between input variables and the target outcome (Pillai, 2024). The model predicts outcomes using a weighted sum of input features, mathematically expressed as:

$$y = w_1x_1 + w_2x_2 + \ldots + w_nx_n + b$$

where y represents the predicted outcome, xi are the input features, wi are the learned model coefficients, and b is the bias term. Since each feature weight directly corresponds to its contribution to the prediction, interpretability is inherent—users can observe how much each feature impacts the final decision. This property makes linear regression widely used in finance (e.g., credit scoring) and economics (e.g., demand forecasting), where transparency is critical (Kumar Ghosh, 2024). However, its primary limitation is the assumption of linearity, which does not always hold in real-world data. Many real-world problems involve non-linear relationships, making linear regression less effective in complex scenarios.

Logistic regression extends linear regression for classification tasks by applying a sigmoid function to transform linear outputs into probabilities. Instead of predicting a continuous outcome, it estimates the probability that a given input belongs to a specific class. The probability of class 1 is given by:

$$P(y=1|x) = \frac{1}{1+e^{-(w_1x_1+w_2x_2+\ldots+w_nx_n+b)}}$$

Logistic regression remains highly interpretable because the learned weights indicate how strongly each feature influences classification. It is widely used in medical diagnosis (e.g., determining the likelihood of disease presence based on patient features) and risk assessment (e.g., evaluating fraud risk in financial transactions). However, logistic regression fails when relationships are highly non-linear and often requires transformations or interactions between features to improve performance (Benhamou et al., 2021).

4.1.2. Decision Trees: Rule-Based Hierarchical Interpretability

Decision trees represent a hierarchical, rule-based approach to AI decision-making, mimicking human reasoning. These models break down decisions into if-then conditions, forming a tree structure where each internal node represents a decision rule, and each leaf node corresponds to a final prediction. The transparency of decision trees lies in their ability to provide explicit reasoning paths, allowing users to trace how specific predictions are made (Mendel & Bonissone, 2021). Mathematically, decision trees split data based on information gain, which measures how well a feature separates the data into classes. Information gain is computed using entropy:

$$IG = H(S) - \sum_{i=1}^{m} \frac{|S_i|}{|S|} H(S_i)$$

where H(S) is the entropy of the original dataset, Si are the subsets created after a split, and m is the number of subsets. A higher information gain indicates a better split, leading to more meaningful decision-making. Decision trees are widely used in finance (e.g., credit approval), healthcare (e.g., diagnosing diseases), and customer segmentation (e.g., identifying high-value customers). However, they suffer from overfitting, especially when trees grow too deep, capturing noise instead of general trends. To mitigate this, techniques such as pruning and ensemble methods like Random Forests are used (Adom & Mahmoud, 2024).

4.1.3. Generalized Additive Models (GAMs): Balancing Non-Linearity and Interpretability

Generalized Additive Models (GAMs) introduce flexibility into interpretability by extending linear models to capture non-linear relationships while remaining transparent (Jones & Wrigley, 1995). Instead of assigning a fixed weight to each feature, GAMs allow for individual feature transformations using smooth functions:

$$y = f_1(x_1) + f_2(x_2) + \ldots + f_n(x_n) + b$$

where fi(x) are non-linear functions, such as splines, applied to each feature independently. These functions allow GAMs to model complex interactions while maintaining interpretability, as each feature's contribution can be separately visualized (de Asis López et al., 2024). GAMs are particularly useful in medicine (e.g., predicting patient survival rates while showing how risk factors interact), climate science (e.g., modeling temperature changes over time), and economic forecasting (e.g., analyzing how consumer behavior evolves with income levels). However, they require careful tuning of smoothing parameters, and their interpretability may be limited when feature interactions are strong (Wood et al., 2022).

4.1.4. Rule-Based Models: Explicit Human-Readable Decision Logic

Rule-based models use explicit logical conditions to make predictions, providing a clear, step-by-step explanation of their decision-making process. Unlike decision trees, rule lists enforce a strict sequence of conditions, while rule sets allow for unordered conditions (Dhamma & Barus, 2025). These models are especially useful when decisions must be easily auditable, such as in legal and healthcare domains. A rule-based medical diagnosis model may look like this: If fever AND sore throat → Predict viral infection. If cough AND shortness of breath → Predict pneumonia. This structure provides direct interpretability, as every decision can be traced back to a set of transparent conditions. However, rule-

based models can become too complex if the number of rules grows significantly, making them harder to interpret in large datasets.

4.1.5. k-Nearest Neighbors (k-NN): Instance-Based Interpretability

Unlike traditional models that learn explicit decision boundaries, k-Nearest Neighbors (k-NN) classifies instances based on the closest historical data points. The model makes a prediction by identifying the k most similar instances in the training set and taking a majority vote (for classification) or averaging their values (for regression). Mathematically, the prediction is given by:

$$\hat{y} = \frac{1}{k} \sum_{i=1}^{k} y_i$$

where yi are the labels of the k-nearest neighbors. k-NN is highly interpretable because users can directly inspect the similar past cases that influenced a given prediction. This makes it useful in medical diagnosis, where doctors can compare new cases with past patient records, and in recommendation systems, where similar past user behavior can inform personalized recommendations (Tang & Wang, 2023). However, k-NN suffers from high computational costs when dealing with large datasets, as every new query requires scanning the entire dataset to find the closest neighbors.

Intrinsically interpretable models play a vital role in AI applications where trust, accountability, and regulatory compliance are crucial. Unlike post-hoc explainability techniques, these models offer built-in transparency, enabling users to directly understand the logic behind AI-driven decisions. While linear models provide simplicity and transparency, decision trees and rule-based models enable explicit reasoning, GAMs balance non-linearity with interpretability, and instance-based methods like k-NN provide contextual case-based reasoning. However, each method has trade-offs between flexibility, complexity, and interpretability (Akbulut et al., 2017). As AI systems become increasingly integrated into high-stakes decision-making, leveraging intrinsically interpretable models will be essential to ensure that AI remains accountable, fair, and understandable to human users.

4.2. Post-Hoc Explainability - Feature Attribution Methods: Understanding Model Inputs' Influence

Feature attribution methods are a category of post-hoc explainability techniques designed to quantify the contribution of each input feature to an AI model's decision. These methods help interpret complex black-box models by assigning importance scores to input features, allowing stakeholders to audit AI systems, detect biases, and improve model transparency. Feature attribution techniques, such as SHAP (Shapley Additive Explanations), LIME (Local Interpretable Model-Agnostic Explanations), and Integrated Gradients, provide insights into model behavior by determining which features have the greatest impact on predictions (Ibrahim et al., 2022).

4.2.1. SHAP (SHapley Additive Explanations): A Game-Theoretic Approach

SHAP is a widely used feature attribution method that is grounded in Shapley values, a concept from cooperative game theory. Shapley values are a way of assigning payouts fairly to each player in a cooperative game, based on their individual contributions to the total value created by the team. In the context of machine learning, this principle is applied to distribute a model's prediction fairly among its input features, essentially determining how much each feature contributes to the final outcome (Shi et al., 2023). The power of SHAP lies in its ability to provide a quantitative measure of each feature's influence on the model's prediction, making it a valuable tool for explaining complex AI models and their decision-making processes.

Unlike simple feature importance methods, which might only rank features based on their overall contribution to the model, SHAP satisfies two critical interpretability properties that make it more reliable and trustworthy. The first property is Consistency. This means that if a feature has a larger impact on the model's decision in one instance, it will consistently receive a higher SHAP value (Makumbura et al., 2024). For example, if increasing the value of a feature leads to a larger increase in the model's output, that feature will always be assigned a higher SHAP value than features with a smaller impact. This consistency is crucial for ensuring that the importance scores provided by SHAP are reliable and intuitive, as they mirror how we would expect the model to behave in real-world scenarios.

The second important property is Local Accuracy, which ensures that the explanation generated by SHAP is faithful to the actual model prediction. Specifically, the sum of the SHAP values for all features in a given instance will exactly equal the prediction made by the model for that instance (Biecek & Burzykowski, 2021). This is an important feature of SHAP,

as it guarantees that the explanation is not just an approximation but a true reflection of the model's decision process. For example, in a classification model, if the model predicts that an applicant is likely to be approved for a loan, the SHAP values for the features will sum up to match that prediction. This guarantees that the explanation is complete and accurate, providing a transparent and trustworthy breakdown of how each feature influenced the decision (Seebold et al., 2024).

SHAP works by computing the difference in the model's predictions when specific features are included or excluded from the input data. The key idea is that the contribution of each feature to the final prediction is determined by considering all possible combinations (or "coalitions") of features. In other words, SHAP evaluates the marginal contribution of each feature across all possible sets of features that could be present in the model, and averages these contributions to calculate a fair attribution (Younisse et al., 2022). This approach ensures that all interactions between features are captured and that each feature is attributed an importance value that reflects its true role in the model's decision-making process.

The process of computing SHAP values mathematically involves determining the marginal contribution of each feature to the prediction, which is calculated for every possible subset of features. To formalize this, the Shapley value for a feature i is given by the equation:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} \Big(f(S \cup \{i\}) - f(S) \Big)$$

- where: ϕi represents the SHAP value for feature i,
- S is a subset of features excluding i,
- f(S) denotes the model's prediction using only features in S.
- N represents the set of all features.

This equation guarantees that every feature's contribution is distributed fairly across all possible feature subsets, maintaining local accuracy (the sum of SHAP values matches the model prediction) and consistency (higher impact features get higher attributions).

SHAP for Different Model Types and Setups: SHAP is applicable to a wide range of machine learning models, with specific implementations optimized for different architectures: Tree SHAP: Designed for tree-based models such as XGBoost, LightGBM, and Random Forests (Santos et al., 2024). It leverages the structure of decision trees to compute exact SHAP values efficiently, avoiding the combinatorial explosion of feature subsets. Deep SHAP: Tailored for deep learning models, combining Shapley values with backpropagation techniques to approximate feature attributions efficiently. Kernel SHAP: A model-agnostic approach that approximates SHAP values for any machine learning model by using a linear regression-based weighting system. It works similarly to LIME but satisfies the theoretical guarantees of Shapley values. Linear SHAP: Optimized for linear regression models, leveraging the linear structure to directly compute feature attributions without iterative approximations.

Each of these implementations adapts the core Shapley value principle to suit the model architecture, balancing accuracy and computational efficiency. SHAP can generate both local and global explanations: Local Explanations: SHAP values can be computed for individual predictions, allowing end-users to understand why a specific decision was made (e.g., why a model denied a loan application). Global Explanations: Aggregating SHAP values across multiple samples provides a global understanding of model behavior, revealing which features drive predictions across an entire dataset (Zhang et al., 2023). Additionally, SHAP summary plots, dependence plots, and decision plots provide intuitive visual representations of how features contribute to model outcomes.

SHAP offers a robust and mathematically grounded approach to feature attribution, making it one of the most widely used interpretability methods for understanding AI decision-making. One of its greatest strengths is its mathematical rigor, as it is the only explanation method based on cooperative game theory that adheres to fairness and consistency axioms (Blesch et al., 2023). These axioms ensure that feature attributions remain stable and unbiased, meaning that

features that contribute more to a prediction receive appropriately higher importance values, while those with no impact are assigned zero contribution. This consistency makes SHAP particularly reliable for use in domains where explainability must adhere to strict fairness and accountability standards, such as healthcare, finance, and legal decision-making (Chen et al., 2024).

Another key advantage of SHAP is its versatility, as it can be applied to a broad range of AI models, from simple linear regression models to highly complex deep neural networks. Unlike some explanation techniques that are model-specific, SHAP is flexible enough to be adapted to tree-based models, ensemble methods, deep learning architectures, and even reinforcement learning systems (Makumbura et al., 2024). This wide applicability makes it a valuable tool across various industries, as it can be used to interpret models trained on structured tabular data, text-based natural language processing tasks, and even image classification problems. Its ability to explain virtually any machine learning model gives it a significant edge over many other feature attribution methods, which often have restrictions on the types of models they can be applied to.

SHAP also excels in interpretability through visualization, providing users with intuitive graphical representations of feature attributions. Some of the most widely used visualizations include force plots, which show how individual feature values push a model's prediction higher or lower, and dependence plots, which illustrate how a feature's impact on the prediction changes across different input values (Santos et al., 2024). These visualizations make it easier for users to understand complex feature interactions and identify patterns that may not be apparent from raw numerical feature importance scores. For example, in a credit scoring application, SHAP force plots can visually demonstrate why a specific applicant was approved or denied a loan by illustrating which factors had the strongest influence on the final decision. This interpretability through visualization makes SHAP not only a powerful analytical tool for data scientists but also an accessible explanation method for business stakeholders and domain experts (Chen et al., 2024).

One of the most distinctive advantages of SHAP over other interpretability methods is its ability to provide both local and global explanations. Local explanations allow users to understand why a specific prediction was made for an individual data point, offering granular insights into the decision-making process. This is particularly useful in high-stakes scenarios where AI-driven decisions impact individual users, such as medical diagnosis recommendations or fraud detection alerts (Ibrahim et al., 2022). On the other hand, SHAP's global explanations aggregate feature attributions across an entire dataset, providing a comprehensive understanding of how different features influence predictions on a larger scale. This dual capability makes SHAP a highly effective tool for both individual decision analysis and broader model auditing, allowing organizations to assess AI fairness, bias, and feature dependencies across datasets (Zhang et al., 2023).

However, despite these strengths, SHAP does come with some notable limitations, particularly regarding its computational cost. Since SHAP values are derived by evaluating all possible feature subsets, the method can become extremely resource-intensive, especially for models with high-dimensional input spaces (Chen et al., 2024). This exponential computational complexity makes SHAP impractical for real-time applications or large-scale machine learning models that require rapid decision-making. To mitigate this, approximations such as Kernel SHAP and Deep SHAP have been developed to reduce computation time, but these come at the cost of slight reductions in accuracy and fidelity (Shi et al., 2023).

Another challenge associated with SHAP is approximation errors. While Kernel SHAP and Deep SHAP aim to make the method more computationally feasible, they rely on approximation techniques that may introduce slight deviations in explanations. These deviations are generally small but can be problematic in domains where absolute precision in attribution is required (Seebold et al., 2024). For example, in regulatory settings where AI explanations are subject to legal scrutiny, even minor inconsistencies in feature attributions could raise concerns about the reliability and accountability of model decisions.

SHAP also assumes feature independence, meaning that it evaluates the importance of each feature under the assumption that all other features are independent. However, in many real-world datasets, features exhibit strong correlations, which can lead to attribution inconsistencies (Shi et al., 2023). For instance, in a medical diagnostic AI system, blood pressure and cholesterol levels are often correlated, but SHAP may not fully capture their joint influence on disease prediction due to its assumption of feature independence. This limitation can sometimes lead to misleading interpretations, particularly in complex datasets where feature interactions play a significant role in model predictions.

Despite these challenges, SHAP remains one of the most widely used and theoretically robust feature attribution methods, offering a powerful blend of fairness, versatility, and interpretability (Makumbura et al., 2024). While its computational demands and assumptions about feature independence present limitations, ongoing advancements in

optimization techniques and hybrid interpretability approaches continue to improve SHAP's applicability, making it an indispensable tool in AI explainability.

4.2.2. LIME (Local Interpretable Model-Agnostic Explanations): Perturbation-Based Explanations

LIME approaches explainability by creating a locally interpretable surrogate model that mimics the behavior of the original AI model for a specific prediction. Instead of analyzing the entire model globally, LIME focuses on a single instance at a time, building a simplified model that helps users understand why the AI arrived at that particular decision. This method is particularly useful for complex black-box models, such as deep neural networks and ensemble learning systems, where direct interpretation of the model's decision-making process is not feasible (Hong & Lin, 2024). The core idea behind LIME is perturbation-based learning, where the model's behavior is studied by slightly altering the input and observing how the output changes. By generating a series of modified versions of the original input, LIME effectively probes the model's decision boundary, creating a dataset of synthetic instances that allow it to approximate how different features contribute to the final prediction. This approach enables LIME to construct a locally interpretable linear approximation of the model's complex decision function, making it easier to identify which factors influenced a given decision (Shin, 2023).

How LIME Works: LIME follows a structured four-step process to generate explanations: Perturbation of Input Features: The first step in LIME is to create synthetic variations of the input instance by modifying feature values slightly. These perturbations are small but significant enough to test how the model reacts to changes in input variables. For example, if the model predicts credit approval based on income, employment history, and credit score, LIME would slightly adjust these values to observe how they affect the outcome (Biecek & Burzykowski, 2021). The purpose of this step is to generate a diverse set of nearby instances around the input, allowing LIME to analyze the sensitivity of the model's decision-making process.

Prediction Observation: Once these perturbed inputs are generated, they are fed into the original black-box model to observe how the predictions change in response to slight variations in input features (Cheng et al., 2022). By analyzing how the model behaves under small changes, LIME gains insights into which features are most influential in the decision-making process.

Surrogate Model Construction: After collecting enough perturbed instances and their corresponding predictions, LIME trains a simplified, interpretable model (such as a linear regression model) to approximate the model's decision boundary within the local region around the instance being explained (Peng et al., 2024). This surrogate model is designed to be interpretable, meaning it assigns weights to features in a way that makes intuitive sense to human users. Instead of analyzing the complex non-linear decision-making process of the black-box model, the surrogate model provides a locally valid, linear approximation that helps users understand which factors were most influential.

Feature Importance Extraction: Once the local surrogate model is trained, LIME extracts feature importance values from it. These importance values indicate how much each input feature contributed to the prediction, making it possible to explain why a particular decision was made.

Given an input instance x, LIME constructs a locally weighted linear approximation of the original model's behavior using the following equation:

$$\hat{f}(x) = w_0 + w_1 x_1 + w_2 x_2 + \ldots + w_n x_n$$

- where: f (x) is the locally approximated function,
- wi represents the importance of feature xi,
- n is the number of input features.

This equation is derived using weighted least squares regression, ensuring that the linear approximation fits the model's decision boundary only in the local region where the instance is being analyzed. Unlike SHAP, which considers the global impact of features, LIME only examines a specific prediction, meaning the explanations are instance-specific and may not generalize across the entire dataset.

Unlike methods such as SHAP, which provide global explanations by averaging feature contributions across multiple instances, LIME is purely a local explanation technique. It generates explanations that are only valid for the specific instance being analyzed, meaning that the importance assigned to features may change significantly for different instances (Zafar & Khan, 2021). This local focus makes LIME well-suited for applications where case-by-case explainability is required, such as justifying individual loan approvals, medical diagnoses, or fraud detection alerts. However, this also means that LIME does not provide a holistic understanding of the model's overall behavior, making it less useful for auditing AI systems or detecting systematic biases. By adopting a perturbation-based approach and constructing locally interpretable models, LIME enables AI practitioners and decision-makers to understand why a particular prediction was made, even when working with highly complex black-box models. However, its reliance on random perturbations introduces some variability in explanations, which can affect reproducibility (Awadallah et al., 2022). Nonetheless, LIME remains a widely used and practical tool for AI explainability, particularly in scenarios where fast, instance-specific insights are more valuable than exhaustive, global feature attributions.

One of LIME's most valuable attributes is its model-agnostic nature, which allows it to be applied to virtually any machine learning model, whether it be a deep neural network, a decision tree, or an ensemble-based system. This flexibility makes LIME particularly useful in real-world applications where a variety of AI models are deployed across different domains (Jouis et al., 2023). Unlike explanation techniques that require access to a model's internal structure—such as gradient-based methods that work specifically with deep learning architectures—LIME treats the model as a black box. Instead of relying on internal computations, it probes the model's inputs and outputs to generate explanations, making it widely applicable. This characteristic is particularly beneficial in industries like healthcare, finance, and law, where AI-driven decisions must be justified transparently without requiring modifications to proprietary or sensitive AI models (Urjitha et al., 2025).

Another key advantage of LIME is its ability to produce highly interpretable and user-friendly explanations. Because LIME simplifies a model's decision boundary using a locally interpretable surrogate model, such as linear regression it provides insights that are easily understood by non-technical users. In critical decision-making areas, such as medical diagnosis or financial credit scoring, this interpretability is crucial. For instance, if an AI model predicts that a patient has a high risk of a particular disease, LIME can highlight which specific symptoms or medical indicators contributed most to that prediction (Shin, 2023). Similarly, in financial lending, LIME can reveal whether factors like income level, credit score, or recent transactions were the driving forces behind a loan rejection. By offering clear and comprehensible justifications, LIME enhances trust in AI models, making them more acceptable for deployment in regulatory-sensitive environments where transparency is a legal or ethical requirement (Cheng et al., 2022).

LIME also stands out for its computational efficiency, particularly in comparison to more resource-intensive interpretability methods like SHAP. Unlike SHAP, which requires evaluating all possible feature combinations to compute Shapley values, LIME approximates feature importance using a perturbation-based approach, significantly reducing the computational burden (Zafar & Khan, 2021). This makes LIME ideal for applications that demand rapid decision-making and real-time explanations. In fraud detection systems, for example, where transactions must be classified as legitimate or suspicious in milliseconds, LIME can generate quick justifications for flagged transactions without introducing significant latency. Similarly, in autonomous driving, where split-second decisions about braking, acceleration, and obstacle avoidance must be explained in real time, LIME's lightweight nature ensures that explanations do not slow down critical AI operations. This speed advantage allows AI-driven systems to maintain high responsiveness while still providing necessary transparency (Jouis et al., 2023).

However, despite its strengths, LIME is prone to instability, which can be a significant drawback in situations where consistency and reproducibility are paramount. Because LIME generates explanations by randomly perturbing input data, multiple runs on the same instance can yield different feature importance rankings. This lack of consistency can undermine trust in AI models, especially in high-stakes applications where explanations must remain stable and reliable over time (Peng et al., 2024). For instance, if a healthcare AI system produces different explanations for the same patient's diagnosis on separate evaluations, it could lead to confusion among medical professionals and reduce confidence in the system's reliability. This instability stems from the fact that LIME constructs a synthetic neighborhood around the data point being explained, meaning that small variations in perturbations can lead to different approximations of the model's decision boundary. Because of this, explanations generated by LIME should be interpreted with caution, particularly in environments where explanation consistency is essential.

Another limitation of LIME is that it lacks global interpretability, meaning that while it can effectively explain individual predictions, it does not provide insights into the broader decision-making patterns of a model. While LIME can clarify why a particular loan application was denied or why an AI system classified a medical scan as abnormal, it does not offer an overarching view of the model's behavior across an entire dataset (Urjitha et al., 2025). This makes LIME

unsuitable for tasks like bias detection, fairness analysis, and model auditing, where understanding the model's overall decision trends is crucial. In contrast, methods like SHAP allow for global interpretability by aggregating explanations across many data points, providing a more comprehensive picture of how different factors influence AI predictions. As a result, while LIME is a useful tool for localized, instance-specific interpretability, it often needs to be supplemented with additional methods to gain a holistic understanding of an AI model's decision-making process (Shin, 2023).

4.2.3. Integrated Gradients: Gradient-Based Feature Attribution

Integrated Gradients (IG) is a powerful gradient-based explanation technique specifically designed to interpret deep learning models. Unlike simpler gradient attribution methods that compute feature importance using the model's gradients at a single input instance, IG ensures that feature attributions remain stable and meaningful across the entire activation path leading to a prediction (Bhat et al., 2022). One of the primary limitations of standard gradient-based approaches is their susceptibility to vanishing gradients, where small or near-zero gradients result in misleading attributions. Integrated Gradients overcomes this issue by considering the entire range of activation values from a baseline input to the actual input, providing a more robust and theoretically sound attribution mechanism.

The IG method works by comparing the model's response to a given input with that of a baseline input, which represents the absence of any meaningful information. The choice of baseline depends on the type of data being analyzed. For example, in an image classification task, a baseline could be a completely black image (i.e., all pixel values set to zero), representing the absence of visual content (Adhikari, 2023). In tabular datasets, the baseline might be a row where all feature values are set to their mean or minimum values. By using a baseline, IG ensures that feature importance is relative to the absence of input, rather than an arbitrary reference point.

How Integrated Gradients Work: Integrated Gradients follow a structured process to compute feature attributions: Baseline Selection: The process begins by selecting a baseline input that represents an absence of meaningful information (Adom & Mahmoud, 2024). The choice of baseline is critical because it serves as the reference point against which the actual input will be compared. This ensures that feature importance is measured relative to a starting state of zero influence. Path Integration: Instead of computing the gradient at a single point, IG interpolates between the baseline and the actual input using a continuous path. This means that instead of jumping directly from the baseline to the actual input, a series of intermediate input points is created along this path, with each step gradually introducing feature values from the actual input. This smooth transition allows IG to capture how the model's prediction changes as the input gradually evolves from the baseline to its final value. Gradient Accumulation: For each intermediate input along the path, IG computes the gradient of the model's output with respect to the input features. These gradients indicate how sensitive the model is to small changes in each feature (Greisbach & Klüver, 2022). The gradients are then accumulated across all the interpolated steps, effectively summing up the total contribution of each feature as the input transitions from the baseline to the actual value. This accumulated gradient value serves as the final attribution score for each feature, providing a measure of how much that feature influenced the model's prediction.

Mathematical Formulation of Integrated Gradients

$$IG_i(x) = (x_i - x_i') \times \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha(x - x'))}{\partial x_i} d\alpha$$

- Where xi represents the actual input,
- X'i is the baseline input,
- F(x) is the model's output,
- α represents interpolation steps between the baseline and actual input.

This equation essentially integrates the gradients along the path from the baseline to the actual input, ensuring that feature importance is accumulated over the entire transition rather than at a single snapshot. By averaging these gradients across multiple interpolated inputs, IG provides a more stable and reliable attribution than traditional gradient-based methods.

How Integrated Gradients Works for Different Models: IG is particularly well-suited for deep learning models, where traditional gradient-based explanations often struggle. It can be applied across various architectures, including feedforward neural networks, convolutional neural networks (CNNs), and transformer-based models used in natural language processing (NLP). For Image Classification Models (CNNs): IG is widely used to generate saliency maps that highlight the most important pixels contributing to a model's classification decision. Instead of computing gradients at a single image, IG integrates gradients across a series of interpolated images, ensuring a more stable and comprehensive attribution of visual features (Feldkamp & Strassburger, 2023). This makes IG especially useful for identifying regions of interest in medical imaging, object detection, and facial recognition tasks. For Natural Language Processing (NLP) Models: IG is applied to transformer-based models such as BERT and GPT, where feature attributions must account for word embeddings and contextual dependencies. Instead of directly computing word-level importance scores, IG accumulates gradients across multiple word representations, capturing how the model attends to different tokens. This approach is particularly valuable in tasks such as sentiment analysis, text classification, and machine translation, where interpretability is essential for debugging and improving model trustworthiness (Naveed et al., 2024).

For Structured Data Models (Tabular Data): In models trained on tabular datasets (such as gradient-boosted decision trees or deep learning models for fraud detection), IG helps identify which numerical or categorical features most significantly contributed to a prediction (Adhikari, 2023). By integrating gradients over interpolated feature values, IG can reveal complex dependencies that simple correlation-based techniques might overlook.

Strengths of Integrated Gradients: One of the biggest advantages of Integrated Gradients is that it avoids the vanishing gradient problem commonly encountered in deep networks. Since IG computes gradients across an entire range of input values rather than at a single snapshot, it provides a more stable and reliable feature attribution than traditional methods (Adom & Mahmoud, 2024). This is particularly beneficial in deep learning models, where direct gradient-based explanations often fail due to disappearing or exploding gradients. Another key strength is that IG preserves two essential properties of interpretability: Sensitivity: If a feature has a direct effect on the model's prediction, IG ensures that its attribution is nonzero, correctly reflecting its contribution. Implementation Invariance: Two models that produce the same output for the same input will always yield identical IG attributions, ensuring consistency across model implementations (Naveed et al., 2024). Additionally, IG is computationally more efficient than methods like SHAP, particularly for deep learning models. Since it only requires computing gradients along a predetermined path, IG reduces the need for exhaustive feature subset evaluations, making it more scalable for large-scale applications.

Limitations of Integrated Gradients: Despite its strengths, Integrated Gradients comes with some challenges. One of the biggest limitations is the selection of the baseline input. The choice of baseline significantly impacts the resulting feature attributions. A poorly chosen baseline (e.g., using all-zero values for a dataset where zero is a meaningful value) can lead to misleading explanations. Selecting an appropriate baseline is particularly difficult for text and categorical data, where defining an "absence of information" is not always straightforward. Another limitation is that IG assumes a linear transition between the baseline and the actual input (Bhat et al., 2022). However, in real-world scenarios, feature interactions are often nonlinear and complex, meaning that interpolating between a baseline and input might not fully capture the feature dependencies. This can sometimes lead to oversimplified attributions that do not account for deeper structural relationships within the data. Additionally, while IG is computationally more efficient than SHAP, it still requires multiple forward and backward passes through the model to compute gradients at different interpolation steps (Naveed et al., 2024). This makes it computationally expensive compared to lightweight attribution methods such as simple gradient-based saliency maps or LIME.

Integrated Gradients is a powerful and theoretically sound technique for interpreting deep learning models, addressing many of the shortcomings of traditional gradient-based explanations. By integrating gradients over a continuous range of inputs, it provides more stable, reliable, and mathematically consistent attributions. However, its reliance on baseline selection and assumptions of linear feature contributions can sometimes limit its accuracy, particularly in models with complex feature interactions (Adom & Mahmoud, 2024). Despite these challenges, IG remains one of the most effective and widely used interpretability methods in deep learning, particularly in image classification, NLP, and structured data analysis, where understanding feature importance is crucial for building trustworthy AI systems.

4.3. Visualization Techniques: Making AI Decisions Interpretable Through Visual Representations

Visualization techniques play a crucial role in enhancing the interpretability of AI models by providing human-readable insights into how a model processes and prioritizes information. These methods convert complex numerical outputs into graphical representations, allowing researchers, practitioners, and end-users to see what parts of an input contribute most to a model's decision (Adadi & Berrada, 2018). Visualization-based explanations are particularly valuable in computer vision, natural language processing (NLP), and structured data analysis, where understanding

how AI models arrive at predictions is critical for debugging, validation, and trust-building. Unlike feature attribution techniques that quantify importance scores, visualization techniques generate intuitive representations of model behavior, often in the form of saliency maps or heatmaps (Arrieta et al., 2020). These methods allow domain experts such as doctors analyzing medical AI decisions or financial analysts reviewing fraud detection models to visually verify whether AI systems focus on relevant and meaningful features when making predictions. While effective, visualization methods are not without limitations, as they may vary in stability and lack consistency across similar inputs.

4.3.1. Saliency Maps: Highlighting Important Input Regions

Saliency maps are one of the most widely used visualization techniques, designed to highlight the most influential regions of an input that contribute to a model's decision. The core idea behind saliency maps is to assign an importance score to each part of the input, whether it's a pixel in an image, a word in a sentence, or a numerical feature in a structured dataset indicating how much impact it had on the final prediction (Bhatt et al., 2020). In computer vision, saliency maps have proven particularly useful in explaining deep learning models, especially convolutional neural networks (CNNs). One of the most popular saliency-based methods is Grad-CAM (Gradient-weighted Class Activation Mapping). Grad-CAM works by computing the gradient of the predicted class score with respect to the feature maps of the last convolutional layer in a CNN. This allows the model to localize the most relevant areas in an image by generating a heatmap overlay, which visually indicates which regions influenced the decision (Lundberg & Lee, 2017). Despite their advantages, saliency maps are not always stable. Small perturbations in the input image such as brightness changes, minor rotations, or cropping can sometimes drastically alter the generated saliency map. This sensitivity issue raises concerns about their reliability, especially in high-stakes domains like healthcare and finance. Additionally, while saliency maps highlight relevant regions, they do not explain the relationships between features, limiting their ability to provide deeper causal insights.

4.3.2. Heatmaps: Color-Coded Feature Importance Representations

Heatmaps extend the concept of visual feature representation beyond computer vision to structured data and NLP models. Unlike saliency maps, which are mainly used in image-based AI, heatmaps are applicable to text-based models, structured numerical datasets, and time-series analysis (Ribeiro et al., 2016). In natural language processing (NLP), heatmaps are widely used in attention-based architectures, such as transformer models (e.g., BERT, GPT-3, and T5). These models rely on an attention mechanism that dynamically assigns different importance scores to input words based on their relevance to the final prediction. Heatmaps provide a graphical representation of attention weights, allowing researchers to visualize which words receive the most focus during model inference. Beyond NLP, heatmaps are also effective in structured data analysis, such as fraud detection and financial modeling. In deep learning models trained on tabular datasets, heatmaps can provide a color-coded visualization of feature importance, showing how different numerical and categorical variables interact in high-dimensional models (Ribeiro et al., 2016). For instance, in a credit risk assessment model, a heatmap could reveal whether the model assigns excessive weight to a borrower's age or location, which could indicate potential bias.

A more advanced technique involves hierarchical clustering of attention heads, which allows researchers to track how different transformer layers contribute to linguistic patterns. This provides a multi-level interpretation of how language models process text, helping linguists and AI researchers better understand deep NLP models. However, while heatmaps offer interpretable insights, they also come with challenges (Vaswani et al., 2017). One limitation is that heatmaps can sometimes misrepresent importance, especially in cases where attention weights are spread thinly across many features. This makes it difficult to pinpoint which individual features truly influenced the prediction (Doshi-Velez & Kim, 2019). Additionally, because heatmaps rely on the underlying attention mechanism, they may not fully capture feature interactions outside of what the model explicitly attends to, potentially missing indirect dependencies.

Table 2 Features between Saliency Maps & Heatmaps on Visualization

Aspect	**Saliency Maps**	**Heatmaps**
Primary Use Case	Computer vision (CNNs)	NLP, structured data, time-series
Visualization Type	Image overlay with highlighted regions	Color-coded importance representation
Interpretability	Shows where the model focused	Shows how much attention was assigned
Algorithmic Approach	Based on gradient activations	Based on attention weights
Common Methods	Grad-CAM, Guided Backpropagation	Transformer-based attention visualization
Limitations	Unstable with small input changes	Can misrepresent true feature impact

4.4. Rule-Based Explanations: Extracting Logical Rules from AI Models

Rule-based explanation techniques translate complex AI decision boundaries into human-readable rules, making AI models more interpretable.

4.4.1. Extracting Human-Readable Rules from Black-Box Models

Recent research has explored how deep learning models can be distilled into symbolic rule sets. For example, CGXplain, a novel approach, extracts concise if-then rules from deep neural networks to approximate decision boundaries. These rule sets help bridge the gap between deep learning's high performance and traditional rule-based AI systems. Rule-based explanations are particularly useful in legal AI and compliance-driven industries, where transparent decision-making is crucial (Gunning & Aha, 2019). However, as models grow in complexity, extracted rules may become too numerous or complex, reducing interpretability.

4.4.2. Surrogate Models: Approximating Black-Box Models with Simplified Interpretable Models

Surrogate models aim to approximate complex AI models with simpler, more interpretable representations. These models retain essential decision-making patterns while being easier to analyze. Types of Surrogate Models: Linear Surrogates: Approximate complex nonlinear relationships with linear functions. Decision Tree Surrogates: Convert black-box model predictions into a tree-based hierarchy (Mueller et al., 2021). Rule-Based Surrogates: Generate symbolic rule sets to approximate decision logic. Surrogate models allow researchers to understand black-box AI behavior in a structured way, but they may fail to fully capture nonlinear dependencies in the original model.

4.4.3. Counterfactual Explanations: Answering "What If" Questions

Counterfactual explanations provide intuitive, actionable insights by describing how an AI decision would change if certain input features were modified. These explanations help users explore alternative scenarios and identify key decision factors. However, generating meaningful counterfactuals in high-dimensional spaces remains a challenge, as some counterfactuals may be mathematically valid but unrealistic in real-world settings. AI interpretability methods provide a rich toolkit for understanding AI decisions, but their effectiveness depends on fidelity, stability, computational efficiency, and human comprehensibility (Confalonieri et al., 2021). Different methods excel in different domains feature attribution techniques help understand black-box models, visualization methods reveal hidden decision patterns, rule-based explanations offer logical reasoning, and counterfactual methods enable actionable insights. However, assessing whether these methods truly enhance AI transparency requires rigorous evaluation metrics. The next section, Evaluating Explainability, explores how interpretability methods are assessed, validated, and benchmarked to ensure trustworthy AI deployment in real-world applications (Miller, 2019).

5. Evaluating Explainability

The evaluation of explainable AI (XAI) systems is an essential process that ensures AI-generated explanations are trustworthy, interpretable, and actionable. As AI continues to make decisions in critical domains such as healthcare, finance, and autonomous systems, evaluating explainability is just as crucial as assessing model accuracy or efficiency (Diakonikolas et al., 2021). Unlike traditional AI model evaluation, which focuses primarily on performance metrics such as accuracy, precision, recall, and F1-score, XAI evaluation requires a multidimensional approach. This approach integrates quantitative performance assessments, cognitive psychology principles, user experience analysis, and domain-specific validation methods. Since the effectiveness of explanations depends on both the AI system's internal decision logic and the human user's ability to understand and act upon the explanations, robust evaluation methodologies are necessary to measure the quality, reliability, and impact of AI explanations (Lipton, 2021).

Several key aspects define the evaluation of explainability, including metrics for assessing explanation quality, usability factors, user trust considerations, and domain-specific validation methods. High-quality explanations should be faithful to the model's decision-making process, complete in covering all influencing factors, and consistent across similar cases (Lundberg et al., 2020). Additionally, explainability should enhance user comprehension without overwhelming the decision-maker with excessive complexity. This section explores the essential methodologies for evaluating the faithfulness, completeness, and consistency of AI explanations, while also addressing human-centered evaluation techniques that measure usability, trust, and decision-making effectiveness (Li et al., 2023).

5.1. Metrics for Evaluating Explanation Quality

One of the primary challenges in evaluating XAI systems is determining whether the provided explanations are faithful, complete, and useful to different stakeholders. The effectiveness of an explanation depends on its fidelity to the model's actual decision process, comprehensiveness in covering influencing factors, and consistency across similar cases.

5.1.1. Fidelity: Measuring the Faithfulness of Explanations

Fidelity refers to the degree to which an explanation accurately represents the true decision-making process of the AI model. A high-fidelity explanation ensures that users are not misled by overly simplified or approximate justifications that do not reflect how the AI system actually arrived at a decision. One of the key issues with many post-hoc explanation methods, such as LIME and SHAP, is that they provide an approximate interpretation rather than a true reflection of the model's internal workings (Molnar et al., 2022). This can lead to explanation discrepancy, where the AI's internal decision-making logic does not perfectly match the explanations, it produces. To assess fidelity, mathematical techniques such as axiomatic attribution methods are used to measure how accurately an explanation aligns with the actual decision process of the AI system. A practical approach to evaluating fidelity involves feature ablation tests. In this process, the most influential features (as identified by the explanation method) are systematically removed from the input, and the AI model's prediction is re-evaluated. If an explanation is truly faithful, removing important features should significantly alter the model's output (Ribeiro et al., 2020). Similarly, counterfactual verification tests how explanations hold up when small input changes occur. If an explanation method is faithful, its attributions should remain stable and reflect genuine model behavior.

5.1.2. Completeness: Assessing Whether an Explanation Covers All Influential Factors

A complete explanation should fully capture all relevant features that contributed to a model's prediction, ensuring that no crucial influencing factors are omitted. However, achieving completeness is challenging because AI models often have complex, high-dimensional decision boundaries, making it difficult to pinpoint all contributing factors. The completeness of an explanation is evaluated by comparing different explanation methods and identifying gaps in feature attributions (Rudin & Radin, 2023). Since no single interpretability technique perfectly captures all relevant model influences, combining multiple approaches such as SHAP for local feature importance, LIME for approximate decision boundaries, and Grad-CAM for visual model attention can provide a more comprehensive assessment of an AI model's reasoning. A completeness scoring mechanism involves testing whether an explanation correctly accounts for all significant decision-driving factors. If an explanation consistently ignores key variables that strongly influence predictions, it lacks completeness (Selvaraju et al., 2020). This is particularly important in high-stakes applications such as medical AI, where failing to recognize a crucial symptom in a diagnosis can led to misleading outcomes.

5.1.3. Consistency: Ensuring Stable and Reproducible Explanations

Consistency is another key factor in evaluating explainability. It ensures that similar input cases receive similar explanations. If an AI system produces highly different justifications for nearly identical inputs, users may lose trust in the explanation system, even if the model itself is performing correctly (Delgado et al., 2022). A critical limitation of many post-hoc explanation techniques is that they lack stability, meaning that running the same explanation method multiple times on similar inputs may yield different explanations. This is particularly problematic in methods like LIME, which uses random perturbations to approximate feature importance resulting in explanations that can vary significantly between runs. To evaluate consistency, explanation methods undergo stochastic consistency analysis, where the variance of an explanation across multiple slightly modified inputs is measured (Yang et al., 2023). If small input changes produce drastically different explanations, the method lacks robustness and should not be used in high-risk AI applications where predictability and stability are paramount.

5.2. Human-Centric Evaluation Methods

While mathematical metrics help assess explanation fidelity, human-centered evaluation methods are crucial for determining whether explanations are actually useful, interpretable, and actionable for different user groups. AI explanations must align with how people process information cognitively, avoiding excessive complexity that overwhelms decision-makers.

5.2.1. User Studies: Assessing Comprehension, Trust, and Decision-Making Effectiveness

User studies play a fundamental role in evaluating XAI from a real-world usability perspective. Comparative studies of different explanation techniques across various domains reveal that user trust, satisfaction, and comprehension vary depending on the task and audience. Findings from such studies highlight that one-size-fits-all explainability approaches often fail, reinforcing the need for customized XAI interfaces based on user expertise and context (Wachter et al., 2022).

A structured user evaluation framework typically involves: Comprehension Tests: Measuring how well users interpret and act upon explanations. Trust Calibration Experiments: Analyzing whether explanations increase or decrease user reliance on AI recommendations. Decision Impact Analysis: Determining whether explanations improve human decision-making outcomes in real-world applications (Zhang et al., 2023).

5.2.2. Cognitive Load Analysis: Balancing Complexity and Simplicity in Explanations

One major issue in XAI evaluation is that overly complex explanations can reduce interpretability instead of enhancing it. Cognitive load theory, a concept from cognitive psychology, suggests that explanations should be informative yet concise to avoid overburdening the user's working memory (Ghassemi et al., 2021). A cognitive load scoring approach can incorporate eye-tracking and response-time analysis to measure how much mental effort a user expends while processing explanations. Research in this area emphasizes designing explanations that are easily digestible while still providing necessary depth.

5.3. Balancing Accuracy and Interpretability in XAI Evaluation

A persistent debate in AI explainability revolves around the trade-off between model accuracy and interpretability. While more complex AI models often outperform simpler interpretable ones, their opacity raises concerns about bias, fairness, and accountability. An alternative approach emphasizes the development of inherently interpretable models rather than relying on post-hoc explanations for black-box models (Hohman et al., 2021). Evidence suggests that, in many cases, interpretable models can achieve accuracy levels comparable to deep learning models, particularly in structured data environments such as finance and healthcare. This indicates that selecting the right model architecture is just as crucial as choosing appropriate explainability methods.

5.4. Assessing Trustworthiness, Usability, and User Satisfaction

The effectiveness of XAI evaluation is incomplete without assessing trustworthiness, usability, and user satisfaction. Trustworthiness: Ensuring That Explanations Align with Human Mental Models. Trust in AI is influenced by multiple factors, including explanation quality, model performance, and psychological user perceptions. A trustworthiness evaluation framework examines whether explanations: Match human expectations and reasoning patterns, are stable and reproducible across similar cases and enhance or degrade user confidence in AI recommendations (Thornton et al., 2022). Usability: Designing User-Friendly Explanation Interfaces. XAI usability have identified key human-computer interaction (HCI) principles for creating intuitive, user-friendly explanation systems. Findings emphasize the importance of: Interactive visualization tools for navigating AI explanations. Customizable explanation depth levels based on user expertise. Minimizing unnecessary complexity to prevent cognitive overload. User Satisfaction: Measuring Perceived Usefulness of XAI Systems. Surveys of user satisfaction in XAI have identified key factors such as explanation clarity, relevance, and actionability (Wang et al., 2021). A quantitative satisfaction metric can be developed, where users rate explanations based on: Transparency: How clear is the explanation? Relevance: Does the explanation address what the user needs? Actionability: Can the user act upon the explanation to make better decisions? Transition to Applications of XAI Across Industries The evaluation of explainable AI systems is a multifaceted challenge requiring a combination of quantitative metrics and human-centered validation methods. While fidelity, completeness, and consistency help assess technical soundness, usability studies, trust analysis, and cognitive load assessments ensure that AI explanations are understandable and actionable in real-world scenarios. As AI continues to be deployed in critical domains, evaluating how XAI methods perform in specific industries is essential (Nauta et al., 2023). The next section, Applications of XAI Across Industries, explores how explainability frameworks are implemented in finance, healthcare, legal systems, and other sectors, shaping the future of responsible AI deployment.

6. Applications of XAI across industries

The integration of Explainable Artificial Intelligence (XAI) across multiple industries has revolutionized decision-making processes, fostering transparency, accountability, and trust in AI-driven systems. As AI technologies become increasingly embedded in critical and high-risk sectors, the demand for interpretable models is growing to ensure fairness, regulatory compliance, and ethical responsibility (Veale & Borgesius, 2023). XAI plays a fundamental role in transforming healthcare, finance, law, and autonomous systems, while also expanding into environmental science, cybersecurity, manufacturing, retail, human resources, and education. These emerging applications demonstrate the versatility of XAI in enhancing decision-making, reducing bias, and ensuring more transparent and effective AI deployments.

XAI in Healthcare: Enhancing Trust in AI-Assisted Medical Decision-Making: The healthcare sector relies on AI models for disease diagnosis, treatment recommendations, and personalized medicine, but the complexity of medical decision-

making necessitates explainability, particularly when human lives are at stake (Jiang et al., 2023). The opacity of deep learning-based diagnostic models has been a major barrier to adoption, as medical practitioners require clear reasoning behind AI-driven diagnoses to validate their accuracy. XAI facilitates clinical decision support systems, allowing physicians to assess the reliability of AI predictions by highlighting the most influential factors in a diagnosis (Buhrmester et al., 2021). In radiology, AI models assist in analyzing medical images for early disease detection. Explainability techniques such as gradient-based class activation mapping provide visual transparency into which regions of an X-ray or MRI scan contributed to a diagnosis. Beyond diagnostics, XAI is revolutionizing personalized medicine by offering interpretability in predictive models that assess treatment efficacy based on patient-specific genetic and clinical data. This level of explainability is crucial for gaining regulatory approval, as interpretable AI models enhance trust between healthcare providers, patients, and medical institutions (Broniatowski et al., 2021).

XAI in Finance: Promoting Fairness and Transparency in Automated Decision Systems: The financial sector extensively employs AI for credit risk assessment, fraud detection, algorithmic trading, and customer service automation. However, AI-driven financial models have historically been criticized for their lack of transparency, leading to concerns about bias, unfair lending practices, and the inability of consumers to challenge AI-based decisions. XAI has become essential for enhancing fairness, regulatory compliance, and trustworthiness in financial decision-making (Lundberg & Lee, 2017). In credit scoring, AI models analyze vast amounts of customer data to determine loan eligibility, but the opacity of black-box models has led to regulatory concerns over potential discrimination. XAI methods allow financial institutions to generate interpretable justifications for credit decisions, enabling customers to understand why they were approved or denied a loan. By highlighting the most influential factors such as income, credit history, or debt-to-income ratio XAI ensures that lending decisions are transparent and actionable. Furthermore, explainable AI assists financial regulators in auditing credit risk models to detect biases, ensuring compliance with anti-discrimination laws. Fraud detection is another critical area where XAI is transforming financial security. Traditional AI models flag suspicious transactions based on learned patterns, but customers and financial analysts often struggle to understand why certain transactions are classified as fraudulent. Counterfactual explanations in XAI allow for contextual insights into fraud detection alerts, identifying specific transaction attributes such as geolocation, spending behavior, or transaction frequency that contributed to the fraud classification (Vaswani et al., 2017). This level of interpretability enables financial institutions to refine fraud detection algorithms while providing customers with actionable insights into securing their accounts. In algorithmic trading, XAI enhances transparency in high-frequency trading strategies, ensuring that financial analysts and regulators can monitor AI-driven investment decisions. By explaining the reasoning behind stock market predictions, XAI enables traders to mitigate risks associated with algorithmic biases or market anomalies.

XAI in Legal Systems: Ensuring Fairness and Accountability in AI-Driven Judicial Processes: The legal sector is increasingly integrating AI for case analysis, sentencing predictions, risk assessment, and legal document processing (Adadi & Berrada, 2018). However, the application of AI in judicial systems has been controversial due to concerns over bias, lack of transparency, and potential injustices arising from opaque decision-making models. XAI plays a pivotal role in ensuring that legal AI systems adhere to principles of fairness, accountability, and due process. One significant application of XAI in the legal domain is recidivism prediction, where AI models estimate the likelihood of a defendant reoffending. Transparent AI frameworks provide interpretable risk assessment reports, explaining which factors such as criminal history, socioeconomic status, or behavioral patterns—contributed to the prediction. This level of explainability enables judges, parole boards, and attorneys to assess AI-driven risk scores critically, ensuring that sentencing and parole decisions are based on fair and unbiased evaluations (Arrieta et al., 2020). Another crucial area is AI-assisted legal research, where natural language processing (NLP) models analyze legal documents, case precedents, and statutory laws. XAI methods allow lawyers and judges to understand which legal references an AI system considers most relevant to a case, ensuring that AI-driven legal interpretations align with established judicial principles. Additionally, bias detection techniques in XAI help legal professionals uncover potential discriminatory patterns in AI-driven sentencing recommendations, reinforcing ethical legal practices (Bhatt et al., 2020).

XAI in Autonomous Systems: Autonomous systems including self-driving cars, robotics, and smart city infrastructures heavily rely on AI to make real-time decisions. However, ensuring that these systems operate safely and transparently is a major challenge. The deployment of XAI in autonomous systems is critical for public trust, regulatory approval, and operational safety (Lundberg & Lee, 2017). In self-driving vehicles, AI models process sensor data to make real-time decisions, such as when to brake, accelerate, or change lanes. XAI techniques provide transparency into why a self-driving car chose a particular action, allowing regulators, engineers, and passengers to validate the reasoning behind AI-driven decisions. Feature importance analysis, combined with trajectory prediction explanations, enables autonomous vehicle systems to communicate how external conditions such as pedestrian movement, traffic signals, or road obstructions—impact AI-driven decisions. In the field of robotics, XAI enhances human-robot interaction by ensuring that automated systems provide understandable justifications for their actions. This is particularly important

in industrial and healthcare robotics, where collaborative robots (cobots) work alongside humans in manufacturing, logistics, and medical environments (Ribeiro, Singh, & Guestrin, 2016). Transparent AI decision-making in robotics ensures that humans can predict, trust, and intervene in robotic actions when necessary, minimizing safety risks. Beyond individual autonomous machines, XAI is shaping smart city infrastructures, where AI-driven traffic management systems optimize urban transportation networks. Explainable AI models provide clear justifications for route recommendations, traffic light adjustments, and congestion predictions, allowing city planners to make data-driven infrastructure improvements while maintaining transparency in AI-driven policy decisions (Vaswani et al., 2017).

XAI in Cybersecurity: Enhancing AI-Based Threat Detection and Risk Analysis: The growing complexity of cyber threats has led organizations to adopt AI-powered security systems for intrusion detection, fraud prevention, and malware analysis. However, many cybersecurity AI models operate as black boxes, making it difficult for security analysts to understand how threats are detected and classified. XAI introduces transparency in cybersecurity by providing interpretable threat detection mechanisms, enabling security teams to validate, refine, and trust AI-driven security recommendations (Doshi-Velez & Kim, 2019). For example, XAI-powered intrusion detection systems (IDS) analyze network traffic to identify anomalous behavior indicative of cyberattacks. Feature attribution techniques highlight which network activity patterns trigger security alerts, allowing cybersecurity teams to determine whether an alert is a false positive or a genuine threat. Additionally, explainable AI in fraud detection enables financial institutions to distinguish between legitimate customer transactions and fraudulent activities, reducing the risk of unnecessary account restrictions (Gunning & Aha, 2019).

XAI in Manufacturing: Improving AI-Driven Quality Control and Predictive Maintenance: AI is playing an increasingly significant role in manufacturing and industrial automation, optimizing processes, predicting equipment failures, and ensuring quality control (Mueller et al., 2021). However, manufacturers need to understand how AI models reach decisions to reduce risks, prevent production defects, and improve efficiency. XAI methods provide transparency in predictive maintenance models, allowing engineers to interpret which machine components are likely to fail and why. This enhances maintenance planning, reduces downtime, and minimizes operational costs. In quality control, XAI techniques highlight which product features or defects influence AI-driven quality assessments, enabling manufacturers to fine-tune inspection processes and improve production consistency (Confalonieri et al., 2021).

XAI in Retail and E-Commerce: Driving Personalized Recommendations and Customer Trust: AI-driven personalization in retail and e-commerce tailor's product recommendations, optimizes pricing strategies, and predicts customer preferences. However, consumers often express concerns about how AI algorithms determine what products they see, particularly in targeted advertising. XAI enhances consumer trust in AI-driven recommendations by explaining why certain products are suggested based on browsing history, purchase behavior, and demographic data (Miller, 2019). In dynamic pricing models, explainability ensures that customers understand why prices fluctuate, reducing concerns about unfair pricing or algorithmic bias. Additionally, XAI in supply chain optimization improves demand forecasting and inventory management by providing interpretable insights into logistics patterns, supplier risks, and stock allocation strategies (Diakonikolas et al., 2021).

XAI in Human Resources: Addressing Bias in AI-Based Hiring and Employee Management: AI-powered hiring and recruitment tools screen candidates, evaluate resumes, and conduct sentiment analysis in interviews. However, concerns over algorithmic bias and discrimination have led to increased scrutiny of AI-based hiring decisions. XAI helps organizations ensure that AI-driven talent acquisition systems provide fair and unbiased candidate assessments by explaining which factors contribute to hiring decisions (Lipton, 2021). For instance, explainability in resume screening AI models ensures that hiring decisions are based on relevant skills rather than indirect demographic biases. Additionally, XAI-powered employee performance evaluation systems provide transparency in promotion, appraisal, and talent retention decisions, ensuring fair career growth opportunities.

XAI in Education: Personalized Learning and Transparent AI-Driven Assessments: AI is transforming education through adaptive learning platforms, automated grading systems, and student performance predictions. However, students and educators need to understand how AI recommendations are generated to ensure transparency and fairness (Lundberg et al., 2020). XAI-driven personalized learning platforms adapt coursework based on student performance, but explainability is crucial for educators to understand why certain learning paths are suggested. In AI-based grading systems, XAI ensures transparency in automated assessments, allowing educators and students to review how assignments are evaluated and graded. Furthermore, XAI can assist in early intervention strategies for at-risk students, highlighting factors that may indicate academic struggles or dropout risks, enabling proactive support (Li et al., 2023).

XAI is driving transformation across diverse industries, ensuring that AI-driven decisions are transparent, interpretable, and trustworthy. Beyond traditional sectors like healthcare, finance, law, and autonomous systems, XAI is expanding into cybersecurity, manufacturing, retail, human resources, and education, enabling fair and ethical AI adoption (Molnar, Casalicchio, & Bischl, 2022). As industries increasingly adopt AI-driven decision-making, the role of XAI in regulatory compliance, bias mitigation, and ethical AI governance will become even more crucial, shaping the future of responsible AI deployment across all sectors.

7. Ethical and Societal Implications of XAI

The rise of Explainable Artificial Intelligence (XAI) marks a significant shift toward transparency and accountability in AI-driven decision-making. As AI systems increasingly influence various aspects of society—including healthcare, finance, law, employment, and governance understanding the ethical and societal ramifications of these technologies has become essential (Ribeiro, Singh, & Guestrin, 2020). While XAI offers a path toward responsible AI deployment, it also presents complex challenges in bias detection, fairness, regulatory compliance, accountability, and security. Ensuring that AI models remain interpretable, equitable, and ethically sound requires interdisciplinary collaboration among technologists, ethicists, policymakers, and domain experts.

7.1. Bias Detection and Mitigation: Addressing Algorithmic Fairness

AI systems, despite their data-driven nature, often perpetuate and even amplify societal biases. Bias arises when historical inequalities, human prejudices, or imbalanced datasets influence model training, leading to unfair outcomes in hiring, lending, law enforcement, and healthcare. XAI provides a powerful toolkit for uncovering hidden biases, enabling organizations to identify, diagnose, and mitigate discriminatory patterns within AI models (Rudin & Radin, 2023). Bias detection in AI requires a granular examination of how features contribute to model predictions. Feature attribution techniques, such as SHAP (Shapley Additive Explanations), highlight which input features disproportionately impact predictions across demographic groups. By analyzing these attributions, developers can determine whether factors such as gender, race, or socioeconomic status unfairly influence decisions. Once biases are detected, XAI allows for real-time adjustments, ensuring that models prioritize fair and unbiased decision-making without compromising performance. Beyond feature attribution, XAI methods such as counterfactual explanations play a crucial role in bias mitigation. Counterfactual reasoning explores how modifying certain inputs could alter a model's decision, revealing whether specific groups experience systematic disadvantages in AI-driven outcomes (Selvaraju et al., 2020). By identifying unfair decision boundaries, organizations can implement corrective measures that align AI models with ethical fairness principles. The effectiveness of bias mitigation in XAI is also dependent on continuous monitoring and re-evaluation of AI systems. Ethical AI frameworks emphasize the need for periodic audits and fairness testing, ensuring that AI models remain accountable and adaptable to societal expectations. Without a structured approach to fairness, XAI may inadvertently provide superficial explanations that fail to address deeper systemic biases (Yang et al., 2023).

7.2. Fairness in Automated Decision-Making: Ensuring Just and Equitable AI Systems

AI is increasingly tasked with high-stakes decision-making, influencing individuals' access to education, employment, credit, and legal rights. Ensuring fairness in these automated decisions is critical to prevent discrimination and uphold ethical AI practices. XAI serves as a bridge between abstract fairness principles and real-world AI implementation, translating complex model behavior into interpretable fairness assessments. Fairness in AI is multidimensional, encompassing concepts such as demographic parity, equal opportunity, and individual fairness. XAI enables stakeholders to analyze whether AI models systematically disadvantage certain groups by providing detailed insights into model decision patterns (Delgado, Barocas, & Levy, 2022). For instance, decision trees and feature attribution models can highlight whether specific demographic characteristics disproportionately affect hiring decisions, ensuring that AI-driven recruitment remains inclusive and unbiased. Additionally, XAI promotes fairness by offering actionable transparency to affected individuals. In financial services, for example, customers receiving AI-driven credit denials can benefit from explanations that outline how financial behaviors influence their eligibility. By understanding how specific attributes impact decisions, individuals are empowered to take corrective actions, promoting agency and fairness in AI interactions. One of the fundamental challenges in ensuring fairness is balancing explainability with algorithmic complexity. AI systems that optimize for fairness often sacrifice predictive accuracy, leading to trade-offs between interpretability and model performance (Wachter, Mittelstadt, & Russell, 2022). XAI frameworks must navigate these competing priorities, ensuring that fairness enhancements do not compromise the reliability and effectiveness of AI systems.

7.3. Regulatory Compliance: Aligning XAI with Legal and Ethical Standards

Regulatory bodies worldwide are increasingly enforcing transparency and accountability requirements for AI systems, recognizing that opaque decision-making models pose risks to individuals' rights and freedoms. XAI plays a pivotal role in ensuring compliance with global AI regulations, particularly in domains where automated decisions significantly impact human lives. One of the most well-known legal frameworks advocating for XAI is the General Data Protection Regulation (GDPR), which mandates a "right to explanation" for individuals affected by automated decision-making. AI-driven decisions, particularly in finance, hiring, and law enforcement, must be accompanied by meaningful, interpretable justifications, allowing individuals to challenge unfair or erroneous outcomes (Zhang et al., 2023). XAI techniques such as decision trees, rule-based models, and interpretable neural networks facilitate GDPR compliance by ensuring that AI-generated explanations are understandable and actionable. Beyond GDPR, emerging AI regulations including the proposed EU AI Act and various national AI ethics guidelines are introducing stringent transparency requirements. High-risk AI systems, particularly those used in healthcare, public safety, and financial services, are required to demonstrate clear and interpretable decision-making processes. XAI assists organizations in meeting these regulatory standards by providing audit trails, fairness reports, and model documentation, ensuring that AI-driven decisions remain accountable and reviewable (Ghassemi, Oakden-Rayner, & Beam, 2021). The role of regulatory-compliant XAI frameworks extends beyond legal adherence; it also fosters trust between organizations and users. As AI systems become more embedded in public services and private enterprises, demonstrating transparent and ethical AI practices becomes a competitive advantage, enhancing user confidence and promoting wider AI adoption.

7.4. Accountability in High-Stakes Domains: Strengthening Oversight and Human-AI Collaboration

AI applications in healthcare, criminal justice, finance, and autonomous systems introduce high-stakes decision-making, where errors or biases can lead to severe consequences. Establishing clear lines of accountability ensures that AI systems remain subject to human oversight, reinforcing ethical AI deployment. In healthcare, for instance, AI-driven clinical decision support systems must provide context-aware explanations that align with medical reasoning (Hohman et al., 2021). XAI frameworks ensure that physicians understand AI-generated diagnoses, allowing them to validate and contextualize recommendations within the broader scope of patient care. Explainability also fosters trust in AI-assisted diagnostics, encouraging greater acceptance of AI-driven medical technologies. Similarly, in criminal justice, predictive AI models influence risk assessment, sentencing recommendations, and parole decisions. The use of black-box AI in judicial proceedings has raised ethical concerns regarding transparency, fairness, and due process. XAI serves as a critical accountability mechanism, ensuring that judicial AI systems provide interpretable justifications for sentencing predictions, preventing bias-driven sentencing disparities (Thornton, Knowles, & Blair, 2022). Financial decision-making is another domain where AI accountability is paramount. AI-driven investment and lending models must remain explainable to financial analysts, regulators, and consumers, ensuring that decisions regarding loans, credit scores, and investment portfolios are transparent and justifiable. A key challenge in AI accountability is balancing automation with human oversight. Over-reliance on AI can lead to automation bias, where human decision-makers blindly trust AI outputs without critical examination. XAI mitigates this risk by promoting human-AI collaboration, ensuring that AI-driven recommendations remain subject to expert validation and ethical considerations (Wang et al., 2021).

7.5. Security and Robustness in XAI: Preventing Manipulation and Misuse

While XAI enhances transparency, it also introduces potential security risks, as adversarial actors may exploit explainability techniques to reverse-engineer model behavior. Malicious attacks on XAI systems can alter feature importance rankings, manipulate counterfactual explanations, or generate misleading attributions, undermining AI integrity (Nauta et al., 2023). Ensuring robust and secure XAI frameworks requires the development of adversarial resilient explanation techniques. AI security research emphasizes the need for robustness verification methods, ensuring that XAI outputs remain consistent and resistant to adversarial perturbations. Additionally, XAI ethics must consider the unintended consequences of AI explanations, ensuring that transparency does not compromise user privacy or proprietary model information. Striking the right balance between explainability and security is critical for safeguarding AI systems against manipulation while maintaining public trust in AI-driven technologies (Broniatowski et al., 2021). The ethical and societal implications of XAI extend far beyond technical AI development, influencing fairness, accountability, regulatory compliance, and security. While explainable AI provides a path toward responsible AI adoption, addressing bias, fairness, and security challenges remains an ongoing effort. The future of XAI requires continuous collaboration between AI researchers, policymakers, and industry leaders, ensuring that AI-driven decision-making remains transparent, fair, and aligned with human values.

8. Challenges and Limitations of XAI

Explainable Artificial Intelligence (XAI) has made significant progress in addressing the transparency and interpretability of AI models, yet several fundamental challenges and limitations persist. While explainability techniques have improved the interpretability of machine learning and deep learning models, key issues remain unresolved, particularly in high-dimensional, complex, and dynamic AI systems. These challenges range from scalability constraints, trade-offs between accuracy and interpretability, lack of standardization in explanation methods, difficulties in handling non-linear relationships and temporal dependencies, ethical concerns regarding manipulations of explanations, and privacy risks (Jiang et al., 2023). Overcoming these limitations is essential for ensuring the reliable and responsible deployment of AI across industries.

8.1. Scalability Issues in XAI for Complex Models

A major challenge in XAI lies in scalability, particularly when applied to deep learning models with millions of parameters and intricate connections. As AI architectures become increasingly complex such as deep neural networks (DNNs), transformers, and ensemble learning techniques generating meaningful explanations becomes computationally expensive and less reliable. Many XAI techniques that work effectively for simple models struggle when extended to deep neural networks due to the intricate, non-linear nature of high-dimensional feature spaces (Lee, Byeon, & Kim, 2022). Moreover, deep learning models operate as black boxes, where decisions result from multiple transformations across hierarchical layers. Traditional feature attribution methods often fail to trace the exact reasoning behind AI predictions, particularly in models that rely on convolutions, recurrent layers, or attention mechanisms. The challenge is to develop scalable XAI approaches that maintain efficiency while providing accurate and interpretable explanations, ensuring that deep learning models remain transparent without degrading computational performance (Sun et al., 2024).

8.2. Trade-off Between Model Accuracy and Interpretability

A persistent dilemma in AI development is the trade-off between model accuracy and interpretability. Highly interpretable models, such as decision trees and linear regression, often lack the predictive power of complex, deep learning-based architectures (Neves et al., 2023). Conversely, more sophisticated models, such as deep neural networks and ensemble methods, offer higher accuracy but at the cost of decreased explainability. This trade-off is particularly critical in high-risk domains, such as healthcare, finance, and criminal justice, where both accuracy and transparency are required. Decision-makers must balance interpretability and predictive performance, ensuring that explanations do not oversimplify complex relationships or introduce misleading simplifications (Knof, Boerger, & Tcholtchev, 2024). Developing intrinsically interpretable models that maintain high accuracy remains one of the most pressing challenges in XAI.

8.3. Lack of Standardization Across Explanation Methods

The absence of universal standards for evaluating XAI techniques complicates the adoption of explainability methods across industries. Currently, there exists a diverse range of explanation techniques, including post-hoc explanations (e.g., SHAP, LIME), intrinsic interpretability models, and surrogate models. However, the absence of a standardized framework makes it difficult to compare and validate different explanation techniques systematically (Vale, El-Sharif, & Ali, 2022). Without widely accepted benchmarking criteria and performance metrics, explanations can become inconsistent across models and applications, leading to ambiguity in their reliability. Some methods may provide contradictory explanations for the same AI decision, making it challenging for users to discern which method offers the most accurate representation of the decision-making process. Establishing clear evaluation metrics and standardized guidelines for XAI is essential to ensure cross-comparability, consistency, and trustworthiness (Mota et al., 2024).

8.4. Challenges in Explaining Non-Linear Relationships and High-Dimensional Data

Real-world AI applications frequently involve non-linear relationships and high-dimensional feature spaces, particularly in fields such as computer vision, genomics, and natural language processing. Many machine learning models capture complex interactions between variables, making it difficult to provide intuitive, human-interpretable explanations (Mota et al., 2024). Feature importance techniques often struggle to accurately represent feature interactions, particularly in cases, where multiple features influence predictions non-linearly. Conventional linear approximation methods, such as LIME, may oversimplify relationships, providing misleading interpretations that fail to capture the true complexity of the model's behavior. Similarly, explaining high-dimensional data such as text embeddings in language models or pixel data in image recognition requires advanced interpretability frameworks that go beyond simple feature attribution methods (Retzlaff et al., 2024). One possible approach involves combining partial dependence plots, accumulated local effects, and contrastive explanations to provide a more comprehensive

understanding of feature interactions. However, refining these methods to handle non-linearity without oversimplification remains a major research challenge.

8.5. Difficulties in Explaining Temporal Dependencies in Sequential Data Models

AI models used in natural language processing (NLP), time-series forecasting, and autonomous systems often rely on sequential data, where present predictions depend on past inputs. Models such as Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, and Transformers encode historical dependencies, making their decision-making process inherently difficult to interpret (Narkhede, 2024). Explaining how an AI model incorporates past information into current predictions is a complex challenge in XAI. Traditional feature attribution techniques are inadequate because they analyze single-instance feature importance rather than tracking how temporal dependencies influence sequential decisions. Visualizing attention patterns in transformers and identifying recurrent patterns in RNNs may offer insights into temporal interpretability, but these methods still require refinement to provide intuitive and human-readable explanations (Belaid et al., 2023).

8.6. Personalized Explanations for Diverse User Groups

The effectiveness of AI explanations depends on the audience consuming them. Developers, regulators, domain experts, and end-users require different levels of detail and technical complexity in AI explanations. AI engineers require explanations that highlight model architecture behavior, enabling them to debug and improve performance (Bhagavatula, Ghela, & Tripathy, 2024). Regulators and policymakers need legal and ethical transparency, ensuring that AI decisions align with compliance frameworks. General users require simplified, intuitive explanations that allow them to understand why a decision was made and what actions they can take. One challenge in XAI is developing adaptive, user-centric explanation techniques that cater to different expertise levels without compromising accuracy or clarity (Friesel & Spinczyk, 2022). AI-generated context-aware explanations that dynamically adjust based on user profiles could enhance accessibility but remain a complex research area.

8.7. Ethical Concerns: Manipulation and Adversarial Attacks on Explanations

While XAI enhances transparency, it also introduces the risk of manipulation. Explanations can be crafted to mislead users, allowing AI developers to justify decisions without revealing underlying biases or inconsistencies in the model (Akhai, 2023). Moreover, XAI methods are vulnerable to adversarial attacks, where malicious actors modify inputs to deceive AI explanations while keeping predictions unchanged. This creates the illusion of fairness or correctness, even when underlying biases persist. Developing robust, tamper-proof XAI frameworks that maintain integrity in adversarial environments is crucial for ensuring the trustworthiness of explainable AI systems (Tian et al., 2023).

8.8. Computational Overhead and Performance Constraints

Many explainability techniques, particularly post-hoc model explanations, impose a computational burden that can reduce the efficiency of real-time AI applications. High computational costs arise from the need to compute multiple perturbations, generate counterfactual instances, or analyze deep feature interactions (Kozielski, Sikora, & Wawrowski, 2025). This limitation is particularly relevant in resource-constrained environments, such as edge computing, IoT devices, and autonomous vehicles, where AI systems must operate with minimal latency. Developing lightweight XAI methods that provide real-time, low-cost explanations without degrading performance is essential for deploying explainability techniques in production systems. The development of XAI has made AI systems more transparent, accountable, and interpretable, but numerous challenges remain unsolved. Issues related to scalability, interpretability-accuracy trade-offs, standardization, high-dimensional feature interactions, sequential data explanations, adversarial security, and computational efficiency must be addressed to ensure the long-term reliability of XAI (Jang, Kim, & Yoon, 2023). Advancements in causal inference, hybrid explainability techniques, and user-adaptive explanations may provide a path forward, enabling AI models to deliver accurate, fair, and interpretable decisions across diverse domains. As AI continues to shape critical aspects of society, overcoming these challenges will be fundamental in ensuring that AI remains a responsible, ethical, and trustworthy tool for decision-making.

9. Proposed Improvements and Future Directions in XAI

Explainable Artificial Intelligence (XAI) is at the forefront of AI research, continuously evolving to enhance transparency, interpretability, and accountability in decision-making systems. As AI becomes increasingly integrated into high-stakes domains, such as healthcare, finance, and autonomous systems, the need for reliable, real-time, and adaptive explainability techniques is more critical than ever. Several key advancements are shaping the future of XAI, addressing existing limitations and expanding its applicability across complex, dynamic, and multimodal AI environments (Ramachandran, 2023).

9.1. Real-Time Explainability in Dynamic and IoT-Based AI Systems

As real-time AI systems gain widespread adoption, particularly in Internet of Things (IoT) environments, the challenge of providing instant, meaningful explanations becomes paramount. IoT applications ranging from smart healthcare monitoring to autonomous transportation networks rely on AI-driven decisions that demand low-latency, high-accuracy interpretability frameworks (Shojaeinasab et al., 2024). Traditional post-hoc XAI techniques, such as feature attribution and model approximation, often introduce computational delays, making them unsuitable for real-time applications. To address this challenge, lightweight attention mechanisms and edge computing have been proposed to deliver instantaneous, context-aware explanations at the device level. By decentralizing the computation process, edge-based XAI can provide localized insights without relying on cloud-based processing, significantly reducing latency and security risks (Dağlarli, 2020). These improvements will be crucial in sectors like smart manufacturing, where automated quality control systems require real-time interpretability to prevent costly production errors.

9.2. Integrating XAI with Edge Computing and Federated Learning

The convergence of XAI, edge computing, and federated learning presents a groundbreaking approach to secure, efficient, and decentralized explainability. AI models are increasingly being deployed in distributed environments, where data privacy and computational efficiency are key concerns (Kosov et al., 2024). Edge-based XAI enables devices to process explanations locally, reducing dependence on centralized data centers while ensuring faster and more context-sensitive insights. Federated learning further enhances this approach by allowing AI models to be trained across multiple decentralized nodes without directly sharing sensitive user data. However, federated learning introduces new challenges for explainability, as distributed models operate across heterogeneous data sources. XAI techniques tailored for federated learning must preserve privacy constraints while offering insights into how AI models adapt to different environments and user behaviors (YazdaniBanafsheDaragh & Malek, 2021). The application of XAI in federated learning is particularly relevant in healthcare AI, where privacy-preserving explainability is essential. For instance, in AI-driven diagnostics, models trained on distributed hospital datasets must provide interpretable justifications for medical decisions without exposing patient-specific data. Future research must focus on privacy-aware explanation methods, ensuring that transparency coexists with data confidentiality.

9.3. Standardization of Explanation Protocols for Cross-Domain Interoperability

One of the major hurdles in XAI adoption is the lack of standardized explanation protocols across industries. AI-driven decisions impact multiple sectors such as finance, legal systems, healthcare, and autonomous systems yet explanation methodologies remain inconsistent, leading to difficulties in evaluating, comparing, and validating XAI techniques across applications (Mokdad et al., 2024). To overcome this challenge, industry-wide efforts are underway to establish universal standards for XAI evaluation metrics, data formats, and communication protocols. Standardization would enable AI systems across different domains to generate, interpret, and exchange explanations seamlessly, fostering greater transparency and regulatory compliance. Additionally, structured explanation taxonomies are being developed to classify XAI techniques based on model architecture, application domain, and intended audience (Prasath & Priya, 2024). This classification will guide researchers and practitioners in selecting appropriate interpretability methods for specific use cases, ensuring that explanations remain contextually relevant and effective.

9.4. Advancing Natural Language Processing (NLP) for Human-Centric Explanations

As AI becomes more integrated into everyday interactions, enhancing human-AI communication through NLP-driven explanations has emerged as a priority in XAI research. Many current explainability techniques rely on visual or mathematical representations, making them inaccessible to non-technical users. Natural Language Processing (NLP)-enabled XAI frameworks are bridging this gap by generating human-like, intuitive explanations tailored to different user groups (Kaushik, Pavithra, & Subbulakshmi, 2024). Recent advancements in large language models (LLMs) have enabled dynamic, dialogue-based interactions, allowing users to engage in conversational queries about AI decisions. Rather than receiving static justifications, users can ask follow-up questions to explore alternative decision paths, causal relationships, and confidence levels in AI outputs (Kamath & Liu, 2021). This approach enhances trust and usability, particularly in customer service AI, automated legal reasoning, and personalized healthcare recommendations. Future research will focus on adaptive NLP models capable of tailoring explanations based on user expertise and context (Miró-Nicolau, Jaume-i-Capó, & Moyà-Alcover, 2024). For example, a physician using an AI-assisted diagnosis tool may receive a detailed, technical explanation, while a patient interacting with the same system may receive a simplified, jargon-free summary of the AI's reasoning.

9.5. Multimodal Explainability: Integrating Text, Images, and Sensor Data

As AI systems increasingly rely on multimodal data sources, the need for cross-modal explanations is becoming crucial. Many AI applications process text, images, video, and sensor data simultaneously, yet current XAI methods struggle to generate cohesive explanations across different data formats (Açar, 2022). Future XAI advancements will focus on hybrid models that combine visual, textual, and numerical explanations to provide a unified interpretability framework. For instance, in AI-driven medical imaging, an XAI system could highlight critical regions in an MRI scan, provide a textual justification for the diagnosis, and correlate it with patient history and test results for a comprehensive clinical explanation. Similarly, in autonomous driving systems, AI-driven vehicles rely on sensor fusion from cameras, LiDAR, and GPS data (Thomas et al., 2024). A multimodal XAI framework would allow vehicles to explain their driving decisions by integrating real-time traffic data, visual scene analysis, and rule-based justifications, ensuring greater transparency and regulatory acceptance.

9.6. Security and Robustness: Preventing Manipulation of AI Explanations

While XAI enhances trust and accountability, it also introduces new vulnerabilities, as adversarial actors can manipulate explanations to obfuscate biases or deceive users. Ensuring the integrity of AI-generated explanations is a growing concern, particularly in financial transactions, cybersecurity, and autonomous systems, where explainability must be both accurate and tamper-resistant (Chauhan, Bahad, & Jain, 2024). Future research will emphasize the development of adversarial robust XAI techniques capable of detecting and mitigating manipulated explanations. AI security frameworks will integrate explanation validation layers, ensuring that AI-generated justifications align with the actual decision-making process rather than being selectively optimized for human perception (Gummadi, Napier, & Abdallah, 2024). Additionally, explainability frameworks will incorporate differential privacy mechanisms, ensuring that sensitive model details remain protected while still providing meaningful insights. This is particularly crucial in healthcare AI and financial risk assessment, where explainability must strike a balance between transparency and confidentiality.

9.7. Ethical and Societal Considerations in Future XAI Research

Beyond technical advancements, the ethical implications of XAI remain a key area of future research. As AI systems influence employment decisions, credit allocations, and legal rulings, ensuring that explanations remain free from bias and aligned with ethical AI principles is critical. XAI frameworks will increasingly incorporate fairness-aware explanations, ensuring that marginalized and underrepresented groups receive equal consideration in AI-driven decisions (Pillai, 2024). Additionally, regulatory bodies will play a more active role in auditing AI explanations, enforcing stricter compliance standards to prevent algorithmic discrimination and explainability-washing—a practice where misleading explanations are used to conceal unfair model behavior. Interdisciplinary research combining AI ethics, psychology, and cognitive science will shape the next generation of human-centric explainability approaches, ensuring that AI explanations align with human reasoning and decision-making processes (Ghosh, 2024). The future of XAI lies in real-time, multimodal, and human-centered explanations that seamlessly integrate across distributed AI ecosystems. By leveraging edge computing, federated learning, NLP-driven interactions, and adversarial robust security frameworks, explainability research is set to enhance AI trustworthiness across diverse applications. As AI continues to advance, ensuring that explanations remain transparent, interpretable, and ethically sound will be essential in establishing AI as a responsible and accountable decision-making tool for the future (Benhamou et al., 2021).

10. Conclusion and Practical Recommendations

The field of Explainable Artificial Intelligence (XAI) has emerged as an essential pillar in the advancement of AI systems, ensuring that machine-driven decisions are transparent, interpretable, and accountable. As AI systems become increasingly embedded in critical decision-making processes, the ability to explain AI-driven outcomes has become a fundamental requirement in domains such as healthcare, finance, legal systems, autonomous systems, and cybersecurity (Mendel & Bonissone, 2021). This research has provided a comprehensive exploration of XAI, covering its theoretical foundations, practical applications, challenges, and future directions. One of the most profound insights derived from this study is the transformative impact of XAI in bridging the gap between complex AI algorithms and human comprehension. AI models, particularly deep learning architectures and ensemble models, often operate as black-box systems, making their decision-making logic inaccessible to users. XAI techniques provide an avenue to demystify these models, enabling stakeholders including business leaders, policymakers, technical professionals, and general users to understand how AI reaches its conclusions. The integration of XAI across industries offers unprecedented opportunities for optimizing decision-making, improving accountability, and fostering trust (Adom & Mahmoud, 2024). In healthcare, explainability enhances AI-assisted diagnostics, ensuring that medical professionals can validate model recommendations. In finance, transparent AI models mitigate bias in credit scoring and fraud

detection. In autonomous systems, XAI contributes to the safety and regulatory compliance of self-driving vehicles. Despite these advancements, the successful adoption and scalability of XAI remain hindered by technical, regulatory, and ethical challenges (de Asis López, Roca-Pardiñas, & Ordóñez, 2024).

10.1. Key Challenges Identified in XAI

Despite significant advancements, XAI faces persistent challenges that hinder its widespread adoption and effectiveness. Among these challenges, several key limitations were identified: Trade-Off Between Model Accuracy and Interpretability, Highly accurate models, such as deep learning networks, often lack transparency, while interpretable models tend to compromise predictive performance. Striking a balance between the two remains an open research challenge (Wood, Goude, & Fasiolo, 2022). Lack of Standardization in XAI Methodologies – The absence of a universal framework for evaluating and comparing explanation techniques makes it difficult to assess the quality and reliability of AI-generated explanations (Dhamma & Barus, 2025). Scalability Issues in High-Dimensional and Dynamic Data – Many XAI techniques struggle to provide real-time, meaningful explanations for complex models trained on large, multimodal datasets. Vulnerability to Manipulation – XAI techniques can be exploited through adversarial attacks, where malicious actors alter explanations to mislead stakeholders, undermining AI integrity (Tang & Wang, 2023). Regulatory and Ethical Considerations – As governments introduce AI transparency regulations, organizations must align their XAI implementations with evolving legal and ethical standards. To address these challenges and fully realize the potential of XAI, structured strategies and best practices must be adopted by industry leaders, policymakers, and researchers.

10.2. Practical Recommendations for XAI Implementation

To ensure the effective deployment of XAI, organizations must adopt a systematic and user-centric approach that integrates explainability into the design, evaluation, and governance of AI systems. The following key recommendations outline actionable strategies for achieving scalable, ethical, and effective explainability in AI-driven applications:

10.2.1. Implement Domain-Specific XAI Strategies

XAI solutions must be customized based on the unique requirements of each industry. Different domains require varying levels of interpretability, and a one-size-fits-all approach is not effective. In healthcare, explanations should be aligned with medical reasoning, ensuring that clinicians can validate and cross-check AI-driven diagnoses (Akbulut et al., 2017). In finance, explainability should focus on regulatory compliance, providing clear justifications for AI-generated loan approvals, fraud detection, and risk assessments. In autonomous systems, AI models must be capable of explaining real-time decision-making processes, ensuring safety and adherence to legal requirements (Ibrahim, Ahenkorah, & Ewusi, 2022). Organizations should develop domain-specific explainability frameworks that optimize accuracy, interpretability, and compliance.

10.2.2. Establish Standardized Evaluation Metrics for XAI

The lack of standardization in evaluating explainability is one of the biggest hurdles in XAI adoption. To ensure consistency and reliability, organizations must: Define clear benchmarks for assessing the quality of AI explanations, including fidelity, completeness, stability, and user comprehension. Develop quantitative and qualitative metrics for measuring explanation effectiveness, ensuring that interpretability does not degrade model performance (Shi et al., 2023). Align XAI implementations with international AI transparency guidelines, ensuring that explanations are uniformly assessed across regulatory jurisdictions. Adopting industry-wide standards will enable AI systems to generate trustworthy, meaningful, and verifiable explanations (Makumbura et al., 2024).

10.2.3. Integrate XAI into AI Governance and Compliance

As AI regulations evolve, organizations must ensure that XAI methodologies align with legal and ethical requirements. Compliance frameworks must include: Automated auditing tools that validate AI explanations in accordance with privacy and fairness laws. Transparency reports that provide stakeholders with a clear, structured overview of AI decision-making processes (Biecek & Burzykowski, 2021). Ethical AI oversight committees to monitor bias, discrimination, and risks associated with AI explanations. Embedding XAI into AI governance structures will ensure that AI remains accountable, fair, and legally compliant.

10.2.4. Enhance User-Centric and Adaptive XAI Interfaces

Effective explainability is not just about technical accuracy; it must also be accessible to diverse user groups. Organizations should: Develop multi-level explanations tailored to different audiences, ensuring that both technical and non-technical users can understand AI reasoning (Seebold et al., 2024). Incorporate interactive explanation systems,

such as conversational AI-driven XAI, allowing users to engage in real-time discussions with AI models. Use natural language generation (NLG) models to provide intuitive, human-readable explanations that enhance AI adoption. By ensuring that AI-generated explanations are understandable, relevant, and adaptable, organizations can improve user trust and decision-making confidence (Younisse, Ahmad, & Abu Al-Haija, 2022).

10.2.5. Strengthen XAI Security Against Adversarial Exploitation

As XAI becomes a key component of AI trustworthiness, its susceptibility to adversarial manipulation must be addressed. To enhance explanation security, AI developers should: Implement robust defense mechanisms against adversarial attacks on feature attribution and model interpretability. Develop tamper-proof explanation validation systems that ensure the authenticity of AI-generated justifications (Santos, Guedes, & Sanchez-Gendriz, 2024). Adopt differential privacy techniques that maintain transparency while protecting proprietary model information. Ensuring XAI security and robustness will safeguard AI trustworthiness and prevent the misuse of explainability techniques.

10.3. Future Research Directions in XAI

To advance the frontiers of explainability, future research in XAI must prioritize expanding its scope, efficiency, and reliability while ensuring its seamless integration into evolving AI ecosystems. Several emerging research directions present significant opportunities for enhancing XAI methodologies, beginning with real-time explainability in AI-driven IoT and autonomous systems, where the development of lightweight, low-latency XAI models is essential for efficient operation in dynamic environments. Additionally, the growing need for multimodal explainability necessitates integrating text, image, and numerical data interpretations to provide a more holistic and interpretable AI framework. As AI systems increasingly adopt federated and decentralized architectures, enhancing privacy-preserving XAI solutions for distributed machine learning becomes crucial in maintaining security and trust. Furthermore, cognitive-aware XAI must align with human cognitive processes to improve AI-human collaboration, making explanations more intuitive and actionable. Ethical AI and bias-resistant explanations remain fundamental to ensuring that XAI models proactively address biases and enhance fairness in AI-driven decision-making. By tackling these critical research challenges, XAI will continue evolving towards greater transparency, adaptability, and inclusivity, reinforcing its role at the intersection of AI innovation, ethics, and regulatory compliance. As AI continues to reshape critical decision-making processes across industries, the importance of XAI will only intensify, underscoring the necessity of designing AI systems that are not only powerful and efficient but also explainable, fair, and aligned with human values.

References

[1] Açar, M. (2022). Explainable AI (XAI). Journal of AI, Robotics & Workplace Automation, 1(4), 323. https://doi.org/10.69554/avxp5177

[2] Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access, 6, 52138-52160. https://doi.org/10.1109/ACCESS.2018.2870052

[3] Adhikari, T. (2023). Towards Explainable AI: Interpretable Models and Feature Attribution. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.4376176

[4] Adom, I., & Mahmoud, M. N. (2024). RB-XAI: Relevance-Based Explainable AI for Traffic Detection in Autonomous Systems. SoutheastCon 2024, 1358–1367. https://doi.org/10.1109/southeastcon52093.2024.10500215

[5] Akbulut, Y., Sengur, A., Guo, Y., & Smarandache, F. (2017). NS-k-NN: Neutrosophic Set-Based k-Nearest Neighbors Classifier. Symmetry, 9(9), 179. https://doi.org/10.3390/sym9090179

[6] Akhai, S. (2023). From Black Boxes to Transparent Machines: The Quest for Explainable AI. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.4390887

[7] Akhai, S. (2024). Towards Trustworthy and Reliable AI. Explainable Artificial Intelligence (XAI) in Healthcare, 89–99. https://doi.org/10.1201/9781003426073-7

[8] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82-115. https://doi.org/10.1016/j.inffus.2019.12.012

[9] Arun Sampaul Thomas, G., Muthukaruppasamy, S., Nandha Gopal, J., Sudha, G., & Saravanan, K. (2024). Unleashing the Power of XAI (Explainable Artificial Intelligence). Explainable AI (XAI) for Sustainable Development, 303–316. https://doi.org/10.1201/9781003457176-18

[10] Awadallah, M. S., de Arriba-Pérez, F., Costa-Montenegro, E., Kholief, M., & El-Bendary, N. (2022). Investigation of Local Interpretable Model-Agnostic Explanations (LIME) Framework with Multi-Dialect Arabic Text Sentiment Classification. 2022 32nd International Conference on Computer Theory and Applications (ICCTA), 116–121. https://doi.org/10.1109/iccta58027.2022.10206274

[11] B.R, Prasath., & V, Priya. (2024). Explainable AI (XAI): Interpretable Model Architectures. Recent Trends in Data Analysis and Data Visualization, 72–79. https://doi.org/10.58532/nbennurch207

[12] Belaid, M. K., Bornemann, R., Rabus, M., Krestel, R., & Hüllermeier, E. (2023). Compare-xAI: Toward Unifying Functional Testing Methods for Post-hoc XAI Algorithms into a Multi-dimensional Benchmark. Explainable Artificial Intelligence, 88–109. https://doi.org/10.1007/978-3-031-44067-0_5

[13] Benhamou, E., Ohana, J.-J., Saltiel, D., & Guez, B. (2021). Explainable AI (XAI) Models Applied to Planning in Financial Markets. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3862437

[14] Bhagavatula, A., Ghela, S., & Tripathy, B. K. (2024). Demystifying the Black Box. Explainable, Interpretable, and Transparent AI Systems, 203–225. https://doi.org/10.1201/9781003442509-12

[15] Bhat, A., Assoa, A. S., & Raychowdhury, A. (2022). Gradient Backpropagation based Feature Attribution to Enable Explainable-AI on the Edge. 2022 IFIP/IEEE 30th International Conference on Very Large Scale Integration (VLSI-SoC), 1–6. https://doi.org/10.1109/vlsi-soc54400.2022.9939601

[16] Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., ... & Eckersley, P. (2020). Explainable machine learning in deployment. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 648-657). https://doi.org/10.1145/3351095.3375624

[17] Biecek, P., & Burzykowski, T. (2021). Local Interpretable Model-agnostic Explanations (LIME). Explanatory Model Analysis, 107–123. https://doi.org/10.1201/9780429027192-11

[18] Biecek, P., & Burzykowski, T. (2021). Shapley Additive Explanations (SHAP) for Average Attributions. Explanatory Model Analysis, 95–106. https://doi.org/10.1201/9780429027192-10

[19] Blesch, K., Wright, M. N., & Watson, D. (2023). Unfooling SHAP and SAGE: Knockoff Imputation for Shapley Values. Explainable Artificial Intelligence, 131–146. https://doi.org/10.1007/978-3-031-44064-9_8

[20] Broniatowski, D. A., Przybocki, M. A., Phillips, P. J., Hahn, C. A., Fontana, P. C., Yates, A. N., & Greene, K. (2021). Four principles of explainable artificial intelligence. National Institute of Standards and Technology. https://doi.org/10.6028/NIST.IR.8312

[21] Buhrmester, V., Münch, D., & Arens, M. (2021). Analysis of explainers of black box deep neural networks for computer vision: A survey. Machine Learning and Knowledge Extraction, 3(4), 966-989. https://doi.org/10.3390/make3040048

[22] Chauhan, D., Bahad, P., & Jain, J. K. (2024). Sustainable AI. Explainable AI (XAI) for Sustainable Development, 1–15. https://doi.org/10.1201/9781003457176-1

[23] Chen, S., Liu, J., Chen, C., Xie, S., & Cheng, Z. (2024). Hybrid Explainable Network Intrusion Detection Framework Based on Shapley Additive Explanations. 2024 IEEE International Symposium on Parallel and Distributed Processing with Applications (ISPA), 1197–1202. https://doi.org/10.1109/ispa63168.2024.00160

[24] Cheng, Z., Miao, Y., & Zhang, X. (2022). An Uncertainty Network Working Mechanism Analysis Method based on Local Interpretable Model-agnostic Explanations. 2022 China Automation Congress (CAC), 4000–4004. https://doi.org/10.1109/cac57257.2022.10055744

[25] Confalonieri, R., Coba, L., Wagner, B., & Besold, T. R. (2021). A historical perspective of explainable Artificial Intelligence. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 11(1), e1391. https://doi.org/10.1002/widm.1391

[26] D, L., Tiwari, R. S., Dhanaraj, R. K., & Kadry, S. (2024). Explainable AI (XAI) for Sustainable Development. Chapman and Hall/CRC. https://doi.org/10.1201/9781003457176

[27] Dağlarli, E. (2020). Explainable Artificial Intelligence (xAI) Approaches and Deep Meta-Learning Models. Advances and Applications in Deep Learning. https://doi.org/10.5772/intechopen.92172

[28] de Asis López, F., Roca-Pardiñas, J., & Ordóñez, C. (2024). Regression analysis with spatially-varying coefficients using generalized additive models (GAMs). Chemometrics and Intelligent Laboratory Systems, 255, 105254. https://doi.org/10.1016/j.chemolab.2024.105254

[29] Dhamma, Y. A., & Barus, S. P. (2025). Sentiment Analysis on Google Reviews Using NaÃ¯ve Bayes, K-Nearest Neighbors, and Logistic Regression to Improve Novotel Services. Journal of Applied Informatics and Computing, 9(1), 106–114. https://doi.org/10.30871/jaic.v9i1.8923

[30] Doshi-Velez, F., & Kim, B. (2019). Towards a rigorous science of interpretable machine learning. arXiv preprint. https://doi.org/10.48550/arXiv.1702.08608

[31] Explainable AI (XAI): Bridging the Gap between Machine Learning and Human Understanding. (2024). Resmilitaris, 10(1). https://doi.org/10.48047/resmil.v10i1.19

[32] Feldkamp, N., & Strassburger, S. (2023). From Explainable AI to Explainable Simulation: Using Machine Learning and XAI to understand System Robustness. ACM SIGSIM Conference on Principles of Advanced Discrete Simulation, 96–106. https://doi.org/10.1145/3573900.3591114

[33] Fernando Delgado, Solon Barocas, and Karen Levy. 2022. An Uncommon Task: Participatory Design in Legal AI. Proc. ACM Hum.-Comput. Interact. 6, CSCW1, Article 51 (April 2022), 23 pages. https://doi.org/10.1145/3512898

[34] Friesel, D., & Spinczyk, O. (2022). Black-box models for non-functional properties of AI software systems. Proceedings of the 1st International Conference on AI Engineering: Software Engineering for AI, 170–180. https://doi.org/10.1145/3522664.3528602

[35] Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. The Lancet Digital Health, 3(11), e745-e750. https://doi.org/10.1016/S2589-7500(21)00208-9

[36] Goenka, A., Srivastava, S., & Helen Victoria, A. (2024). Introduction to Deployable AI for Cutting-Edge Technologies. Explainable AI (XAI) for Sustainable Development, 271–285. https://doi.org/10.1201/9781003457176-16

[37] Greisbach, A., & Klüver, C. (2022). Determining Feature Importance in Self-Enforcing Networks to achieve Explainable AI (xAI). Proceedings - 32. Workshop Computational Intelligence: Berlin, 1. - 2. Dezember 2022, 237–256. https://doi.org/10.58895/ksp/1000151141-16

[38] Gunning, D., & Aha, D. W. (2019). DARPA's explainable artificial intelligence program. AI Magazine, 40(2), 44-58. https://doi.org/10.1609/aimag.v40i2.2850

[39] Hohman, F., Kahng, M., Pienta, R., & Chau, D. H. (2021). Visual analytics in deep learning: An interrogative survey for the next frontiers. IEEE Transactions on Visualization and Computer Graphics, 27(2), 1563-1579. https://doi.org/10.1109/TVCG.2018.2843369

[40] Hong, S. Y., & Lin, L. P. (2024). Skin Lesion Classification: A Deep Learning Approach with Local Interpretable Model-Agnostic Explanations (LIME) for Explainable Artificial Intelligence (XAI). JOIV : International Journal on Informatics Visualization, 8(3–2), 1536. https://doi.org/10.62527/joiv.8.3-2.3022

[41] Hughes, R., Edmond, C., Wells, L., Glencross, M., Zhu, L., & Bednarz, T. (2020). eXplainable AI (XAI). SIGGRAPH Asia 2020 Courses. https://doi.org/10.1145/3415263.3419166

[42] Ibrahim, B., Ahenkorah, I., & Ewusi, A. (2022). Explainable Risk Assessment of Rockbolts' Failure in Underground Coal Mines Based on Categorical Gradient Boosting and SHapley Additive exPlanations (SHAP). Sustainability, 14(19), 11843. https://doi.org/10.3390/su141911843

[43] Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. 2021. Robustness meets algorithms. Commun. ACM 64, 5 (May 2021), 107–115. https://doi.org/10.1145/3453935

[44] Jang, H., Kim, S., & Yoon, B. (2023). An eXplainable AI (XAI) model for text-based patent novelty analysis. Expert Systems with Applications, 231, 120839. https://doi.org/10.1016/j.eswa.2023.120839

[45] Jiang, H. H., Brown, L., Cheng, J., Khan, M., Gupta, A., Workman, D., Hanna, A., Flowers, J., & Gebru, T. (2023). AI art and its impact on artists. In Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society (pp. 363–374). Association for Computing Machinery. https://doi.org/10.1145/3600211.3604681

[46] Jones, K., & Wrigley, N. (1995). Generalized Additive Models, Graphical Diagnostics, and Logistic Regression. Geographical Analysis, 27(1), 1–18. Portico. https://doi.org/10.1111/j.1538-4632.1995.tb00333.x

[47] Jouis, G., Mouchère, H., Picarougne, F., & Hardouin, A. (2023). A methodology to compare XAI explanations on natural language processing. Explainable Deep Learning AI, 191–216. https://doi.org/10.1016/b978-0-32-396098-4.00016-8

[48] Kamath, U., & Liu, J. (2021). XAI: Challenges and Future. Explainable Artificial Intelligence: An Introduction to Interpretable Machine Learning, 303–310. https://doi.org/10.1007/978-3-030-83356-5_8

[49] Kaushik, K., Pavithra, L. K., & Subbulakshmi, P. (2024). Applications of XAI in Modern Automotive, Financial, and Manufacturing Sectors. Explainable, Interpretable, and Transparent AI Systems, 31–52. https://doi.org/10.1201/9781003442509-3

[50] Knof, H., Boerger, M., & Tcholtchev, N. (2024). Quantitative Evaluation of xAI Methods for Multivariate Time Series - A Case Study for a CNN-Based MI Detection Model. Explainable Artificial Intelligence, 169–190. https://doi.org/10.1007/978-3-031-63803-9_9

[51] Kosov, P., El Kadhi, N., Zanni-Merk, C., & Gardashova, L. (2024). Advancing XAI: new properties to broaden semantic-based explanations of black-box learning models. Procedia Computer Science, 246, 2292–2301. https://doi.org/10.1016/j.procs.2024.09.560

[52] Kozielski, M., Sikora, M., & Wawrowski, Ł. (2025). Towards consistency of rule-based explainer and black box model — Fusion of rule induction and XAI-based feature importance. Knowledge-Based Systems, 311, 113092. https://doi.org/10.1016/j.knosys.2025.113092

[53] Kumar Ghosh, D. (2024). Perspective Chapter: Linear Regression and Logistic Regression Models. Recent Advances in Biostatistics. https://doi.org/10.5772/intechopen.1003183

[54] Kumari, M., Chaudhary, A., & Narayan, Y. (2022). Explainable AI (XAI): A Survey of Current and Future Opportunities. Explainable Edge AI: A Futuristic Computing Perspective, 53–71. https://doi.org/10.1007/978-3-031-18292-1_4

[55] Lee, D., Byeon, S., & Kim, K. (2022). An Inspection of CNN Model for Citrus Canker Image Classification Based on XAI: Grad-CAM. The Korean Data Analysis Society, 24(6), 2133–2142. https://doi.org/10.37727/jkdas.2022.24.6.2133

[56] Li, X., Xiong, H., Li, X., Zhang, X., Liu, J., Jiang, H., Chen, Z., & Dou, D. (2023). G-LIME: Statistical learning for local interpretations of deep neural networks using global priors. Artificial Intelligence, 314, 103823. https://doi.org/10.1016/j.artint.2022.103823

[57] Lipton, Z. C. (2021). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue, 16(3), 31-57. https://doi.org/10.1145/3236386.3241340

[58] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765-4774. https://doi.org/10.48550/arXiv.1705.07874

[59] Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S. I. (2020). From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, 2(1), 56-67. https://doi.org/10.1038/s42256-019-0138-9

[60] Makumbura, R. K., Mampitiya, L., Rathnayake, N., Meddage, D. P. P., Henna, S., Dang, T. L., Hoshino, Y., & Rathnayake, U. (2024). Advancing water quality assessment and prediction using machine learning models, coupled with explainable artificial intelligence (XAI) techniques like shapley additive explanations (SHAP) for interpreting the black-box nature. Results in Engineering, 23, 102831. https://doi.org/10.1016/j.rineng.2024.102831

[61] Mendel, J. M., & Bonissone, P. P. (2021). Critical Thinking About Explainable AI (XAI) for Rule-Based Fuzzy Systems. IEEE Transactions on Fuzzy Systems, 29(12), 3579–3593. https://doi.org/10.1109/tfuzz.2021.3079503

[62] Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 267, 1-38. https://doi.org/10.1016/j.artint.2018.07.007

[63] Miró-Nicolau, M., Jaume-i-Capó, A., & Moyà-Alcover, G. (2024). Assessing fidelity in XAI post-hoc techniques: A comparative study with ground truth explanations datasets. Artificial Intelligence, 335, 104179. https://doi.org/10.1016/j.artint.2024.104179

[64] Mokdad, S. I., Khalid, A., Nasr, D., & Talib, M. A. (2024). Interpretable deep learning: evaluating YOLO models and XAI techniques for video annotation. IET Conference Proceedings, 2023(39), 487–496. https://doi.org/10.1049/icp.2024.0532

[65] Molnar, C., Casalicchio, G., & Bischl, B. (2022). Interpretable machine learning–A brief history, state-of-the-art and challenges. In Explainable AI in Healthcare (pp. 1-21). Springer, Cham. https://doi.org/10.1007/978-3-030-72381-7_1

[66] Mota, B., Faria, P., Corchado, J., & Ramos, C. (2024). Explainable Artificial Intelligence Applied to Predictive Maintenance: Comparison of Post-Hoc Explainability Techniques. Explainable Artificial Intelligence, 353–364. https://doi.org/10.1007/978-3-031-63803-9_19

[67] Mueller, S. T., Hoffman, R. R., Clancey, W., Emrey, A., & Klein, G. (2021). Explanation in human-AI systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable AI. arXiv preprint. https://doi.org/10.48550/arXiv.1902.01876

[68] Namrita Gummadi, A., Napier, J. C., & Abdallah, M. (2024). XAI-IoT: An Explainable AI Framework for Enhancing Anomaly Detection in IoT Systems. IEEE Access, 12, 71024–71054. https://doi.org/10.1109/access.2024.3402446

[69] Narkhede, J. (2024). Comparative Evaluation of Post-Hoc Explainability Methods in AI: LIME, SHAP, and Grad-CAM. 2024 4th International Conference on Sustainable Expert Systems (ICSES), 826–830. https://doi.org/10.1109/icses63445.2024.10762963

[70] Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., van Keulen, M., & Seifert, C. (2023). From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI. arXiv preprint arXiv:2201.08164. https://doi.org/10.48550/arXiv.2201.08164

[71] Naveed, S., Stevens, G., & Robin-Kern, D. (2024). An Overview of the Empirical Evaluation of Explainable AI (XAI): A Comprehensive Guideline for User-Centered Evaluation in XAI. Applied Sciences, 14(23), 11288. https://doi.org/10.3390/app142311288

[72] Neves, L., Martinez, J., Longo, L., Roberto, G., Tosta, T., de Faria, P., Loyola, A., Cardoso, S., Silva, A., do Nascimento, M., & Rozendo, G. (2023). Classification of H&E Images via CNN Models with XAI Approaches, DeepDream Representations and Multiple Classifiers. Proceedings of the 25th International Conference on Enterprise Information Systems, 354–364. https://doi.org/10.5220/0011839400003467

[73] Palacio, S., Lucieri, A., Munir, M., Ahmed, S., Hees, J., & Dengel, A. (2021). XAI Handbook: Towards a Unified Framework for Explainable AI. 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), 3759–3768. https://doi.org/10.1109/iccvw54120.2021.00420

[74] Peng, L., Lv, S., & Wang, L. (2024). Explainable machine learning techniques based on attention gate recurrent unit and local interpretable model-agnostic explanations for multivariate wind speed forecasting. Journal of Forecasting, 43(6), 2064–2087. Portico. https://doi.org/10.1002/for.3097

[75] Pillai, V. (2024). Enhancing the Transparency of Data and ML Models Using Explainable AI (XAI). SSRN Electronic Journal. https://doi.org/10.2139/ssrn.4991713

[76] Ramachandran, M. (2023). Quality framework for explainable artificial intelligence (XAI) and machine learning applications. Explainable Artificial Intelligence (XAI): Concepts, Enabling Tools, Technologies and Applications, 115–138. https://doi.org/10.1049/pbpc062e_ch7

[77] Retzlaff, C. O., Angerschmid, A., Saranti, A., Schneeberger, D., Röttger, R., Müller, H., & Holzinger, A. (2024). Post-hoc vs ante-hoc explanations: xAI design guidelines for data scientists. Cognitive Systems Research, 86, 101243. https://doi.org/10.1016/j.cogsys.2024.101243

[78] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135-1144. https://doi.org/10.1145/2939672.2939778

[79] Ribeiro, M. T., Singh, S., & Guestrin, C. (2020). Beyond accuracy: Behavioral testing of NLP models with CheckList. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 4902-4912). https://doi.org/10.18653/v1/2020.acl-main.442

[80] Rudin, C., & Radin, J. (2023). Why are we using black box models in AI when we don't need to? A lesson from an explainable AI competition. Harvard Data Science Review, 1(2). https://doi.org/10.1162/99608f92.5a8a3a3d

[81] Santos, M. R., Guedes, A., & Sanchez-Gendriz, I. (2024). SHapley Additive exPlanations (SHAP) for Efficient Feature Selection in Rolling Bearing Fault Diagnosis. Machine Learning and Knowledge Extraction, 6(1), 316–341. https://doi.org/10.3390/make6010016

[82] Seebold, P., Kaye, M. K., Kim, S., & Nam, C. S. (2024). Explainable AI-based Shapley Additive Explanations for Remaining Useful Life Prediction using NASA Turbofan Engine Dataset. 2024 IEEE 3rd International Conference on Computing and Machine Intelligence (ICMI), 1–5. https://doi.org/10.1109/icmi60790.2024.10586061

[83] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2020). Grad-CAM: Visual explanations from deep networks via gradient-based localization. International Journal of Computer Vision, 128(2), 336-359. https://doi.org/10.1007/s11263-019-01228-7

[84] Sharma, D., Koundilya, V., & Verma, S. (2020). Explainable AI(XAI): A Review. International Journal of Psychosocial Rehabilitation, 56498–56502. https://doi.org/10.61841/v24i5/400345

[85] Shi, Y., Cai, Y., Lou, S., & Chen, Y. (2023). Explainable prediction of deposited film thickness in IC fabrication with CatBoost and SHapley Additive exPlanations (SHAP) models. Applied Intelligence, 54(1), 246–263. https://doi.org/10.1007/s10489-023-05121-2

[86] Shin, J. (2023). Feasibility of local interpretable model-agnostic explanations (LIME) algorithm as an effective and interpretable feature selection method: comparative fNIRS study. Biomedical Engineering Letters, 13(4), 689–703. https://doi.org/10.1007/s13534-023-00291-x

[87] Shojaeinasab, A., Jalayer, M., Baniasadi, A., & Najjaran, H. (2024). Unveiling the Black Box: A Unified XAI Framework for Signal-Based Deep Learning Models. Machines, 12(2), 121. https://doi.org/10.3390/machines12020121

[88] Sun, C., Xu, H., Chen, Y., & Zhang, D. (2024). AS-XAI: Self-Supervised Automatic Semantic Interpretation for CNN. Advanced Intelligent Systems, 6(12). Portico. https://doi.org/10.1002/aisy.202470055

[89] Tang, Y., & Wang, C.-L. (2023). SelB-k-NN: A Mini-Batch K-Nearest Neighbors Algorithm on AI Processors. 2023 IEEE International Parallel and Distributed Processing Symposium (IPDPS), 831–841. https://doi.org/10.1109/ipdps54959.2023.00088

[90] Tchuente, D., Lonlac, J., & Kamsu-Foguem, B. (2024). A methodological and theoretical framework for implementing explainable artificial intelligence (XAI) in business applications. Computers in Industry, 155, 104044. https://doi.org/10.1016/j.compind.2023.104044

[91] Thornton, L., Knowles, B., & Blair, G. (2022). The alchemy of trust: The creative act of designing trustworthy socio-technical systems. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22) (pp. 1387–1398). Association for Computing Machinery. https://doi.org/10.1145/3531146.3533196

[92] Tian, Z., Zhang, C., Sood, K., & Yu, S. (2023). Inferring Private Data from AI Models in Metaverse through Black-box Model Inversion Attacks. 2023 IEEE International Conference on Metaverse Computing, Networking and Applications (MetaCom), 49–56. https://doi.org/10.1109/metacom57706.2023.00051

[93] Urjitha, P., Shamanth Showri, N. R., Koushik, S. V., Shreya, C. R., & Divya, C. D. (2025). Enhancing model transparency: Integrating local interpretable model agnostic explanations and SHapley additive exPlanations for explainable artificial intelligence in Juvenile onset diabetes prediction. Data Science & Exploration in Artificial Intelligence, 83–89. https://doi.org/10.1201/9781003589273-13

[94] Vale, D., El-Sharif, A., & Ali, M. (2022). Explainable artificial intelligence (XAI) post-hoc explainability methods: risks and limitations in non-discrimination law. AI and Ethics, 2(4), 815–826. https://doi.org/10.1007/s43681-022-00142-y

[95] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998-6008. https://doi.org/10.48550/arXiv.1706.03762

[96] Veale, M., & Borgesius, F. Z. (2023). Demystifying the Draft EU Artificial Intelligence Act—Analysing the good, the bad, and the unclear elements of the proposed approach. Computer Law Review International, 22(4), 97-112. https://doi.org/10.9785/cri-2021-220402

[97] Wachter, S., Mittelstadt, B., & Russell, C. (2022). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harvard Journal of Law & Technology, 31(2), 841-887. https://doi.org/10.2139/ssrn.3063289

[98] Wang, D., Andres, J., Weisz, J. D., Oduor, E., & Dugan, C. (2021). AutoDS: Towards human-centered automation of data science. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (Article 79). Association for Computing Machinery. https://doi.org/10.1145/3411764.3445526

[99] Wood, S. N., Goude, Y., & Fasiolo, M. (2022). Interpretability in Generalized Additive Models. Interpretability for Industry 4.0 : Statistical and Machine Learning Approaches, 85–123. https://doi.org/10.1007/978-3-031-12402-0_4

[100] Yang, K., Klein, D., Peng, N., & Tian, Y. (2023). DOC: Improving long story coherence with detailed outline control. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 3378–3465). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.190

[101] YazdaniBanafsheDaragh, F., & Malek, S. (2021). Deep GUI: Black-box GUI Input Generation with Deep Learning. 2021 36th IEEE/ACM International Conference on Automated Software Engineering (ASE), 905–916. https://doi.org/10.1109/ase51524.2021.9678778

[102] Younisse, R., Ahmad, A., & Abu Al-Haija, Q. (2022). Explaining Intrusion Detection-Based Convolutional Neural Networks Using Shapley Additive Explanations (SHAP). Big Data and Cognitive Computing, 6(4), 126. https://doi.org/10.3390/bdcc6040126

[103] Zafar, M. R., & Khan, N. (2021). Deterministic Local Interpretable Model-Agnostic Explanations for Stable Explainability. Machine Learning and Knowledge Extraction, 3(3), 525–541. https://doi.org/10.3390/make3030027

[104] Zhang, J., Li, C., Yin, Y. et al. Applications of artificial neural networks in microorganism image analysis: a comprehensive review from conventional multilayer perceptron to popular convolutional neural network and potential visual transformer. Artif Intell Rev 56, 1013–1070 (2023). https://doi.org/10.1007/s10462-022-10192-7

[105] Zhang, S., Lei, H., Zhou, Z., Wang, G., & Qiu, B. (2023). Fatigue life analysis of high-strength bolts based on machine learning method and SHapley Additive exPlanations (SHAP) approach. Structures, 51, 275–287. https://doi.org/10.1016/j.istruc.2023.03.060