Tech-Sphere Journal of Pure and Applied Sciences (TSJPAS)
A Subsidiary of Tech-Sphere Multidisciplinary International Journal (TSMIJ)
Akinsiku. Vol 2, Issue 1, 2025 Publication Edition
ISSN: 3092-9598

Literature Review on Explainable Artificial Intelligence (XAI): Techniques, Tools, and Applications

Research Article https://stem.techspherejournal.com

Author Details
Akinsiku Ayokunle Michael
Computer Science Department, The Federal Polytechnic Ado-Ekiti, Ekiti State, Nigeria.

**Corresponding author's email*: akinsiku_am@fedpolyado.edu.ng
DOI: https://doi.org/10.5281/zenodo.15870683

ABSTRACT

Explainable Artificial Intelligence (XAI) refers to a class of methods and tools that make the decision-making processes of AI systems transparent, understandable, and accountable to human users, particularly in high-stakes applications such as healthcare, finance, autonomous systems, and cybersecurity where opaque models can hinder trust, safety, and compliance. With growing ethical and regulatory concerns around black-box AI models, XAI has become essential for ensuring interpretability, fairness, and responsible AI deployment. This paper presents a comprehensive literature review on XAI by first establishing its conceptual foundation, including definitions, explanation types, and the needs of various stakeholders. It then reviews a wide array of XAI techniques, distinguishing between model-specific and model-agnostic methods, and highlights visualization tools, surrogate models, and rule-based explanations. The study further analyzes prominent XAI libraries and platforms such as InterpretML, AIX360, Captum, and AWS Clarify, using evaluation criteria like fidelity, stability, complexity, and generalizability. Real-world applications across critical domains are discussed to demonstrate the value of XAI in enhancing trust and decision support. Finally, the paper identifies key challenges such as trade-offs between accuracy and interpretability, lack of standards, and explainability in deep models, while proposing future research directions involving causal inference, federated AI, human-centric design, and transparency in large language models and reinforcement learning systems.

Keywords: Explainable Artificial Intelligence (XAI), Model Interpretability, AI Transparency, XAI Applications, Responsible AI.

1 Introduction

Artificial Intelligence (AI), particularly driven by machine learning (ML) and deep learning (DL), has achieved remarkable milestones in domains such as healthcare, finance, transportation, and cybersecurity [1]. These achievements are primarily powered by complex models like deep neural networks, ensemble learning algorithms, and support vector machines, which, despite their predictive accuracy, often operate as "black boxes [2]." This black-box nature implies that the internal workings of these models are typically opaque, making it difficult for users and stakeholders to understand or interpret how a given input leads to a specific output.

In critical sectors where AI-driven decisions have direct implications on human lives, such as medical diagnostics, loan approvals, or criminal justice, the need for transparency becomes not only a technical issue but also a moral and legal imperative. As a result, Explainable Artificial Intelligence (XAI) has emerged as a growing field aimed at making AI models more interpretable, trustworthy, and accountable. The ability to explain how and why AI systems make decisions is fundamental for building user trust, ensuring regulatory compliance like the General Data Protection Regulation

(GDPR), and the Health Insurance Portability and Accountability (HIPAA), facilitating debugging by developers, and promoting the ethical deployment of AI technologies [3].

1.1 Problem Statement

Despite the growing integration of AI across domains, most models lack transparency in their decision-making processes. This opacity poses serious challenges including reduced user trust, difficulties in model validation, potential bias or discrimination, and increased legal and ethical risks. The absence of interpretability mechanisms in AI systems can result in user reluctance to adopt AI tools, misinterpretation of outputs, and limited regulatory acceptance [4]. Therefore, there is a pressing need for systematic exploration of the available techniques and tools that promote explainability in AI systems.

1.2 Objectives of the Review

This literature review aims to:

a) Examine and categorize existing XAI techniques, highlighting both model-specific and model-agnostic approaches.
b) Analyse widely-used XAI tools and frameworks, focusing on their features, use cases, and limitations.
c) Explore practical applications of XAI across various domains including healthcare, finance, cybersecurity, and autonomous systems.
d) Identify current challenges, gaps, and potential future directions in the field of XAI.

By synthesizing current literature, this study seeks to offer a comprehensive understanding of the XAI landscape and provide valuable insights for researchers, developers, and practitioners working toward responsible and interpretable AI systems.

1.3 Scope and Structure

The paper is structured as follows:

a) Section 2 introduces the foundational concepts of XAI, including types of explanations and their significance across different stakeholders.
b) Section 3 presents a thematic review of the major XAI techniques, detailing their methodologies, strengths, and limitations.
c) Section 4 discusses available XAI tools and frameworks, comparing their capabilities and deployment contexts.
d) Section 5 explores the real-world applications of XAI in key sectors, providing examples and case studies.
e) Section 6 outlines the major challenges and limitations faced by the field.
f) Section 7 proposes future research directions and trends.
g) Finally, Section 8 concludes with key takeaways and reflections on the evolving role of XAI in modern AI development.

Figure 1 shows the Research Methodological approach adopted in this study.

Tech-Sphere Journal of Pure and Applied Sciences (TSJPAS)
A Subsidiary of Tech-Sphere Multidisciplinary International Journal (TSMIJ)
Akinsiku. Vol 2, Issue 1, 2025 Publication Edition
ISSN: 3092-9598

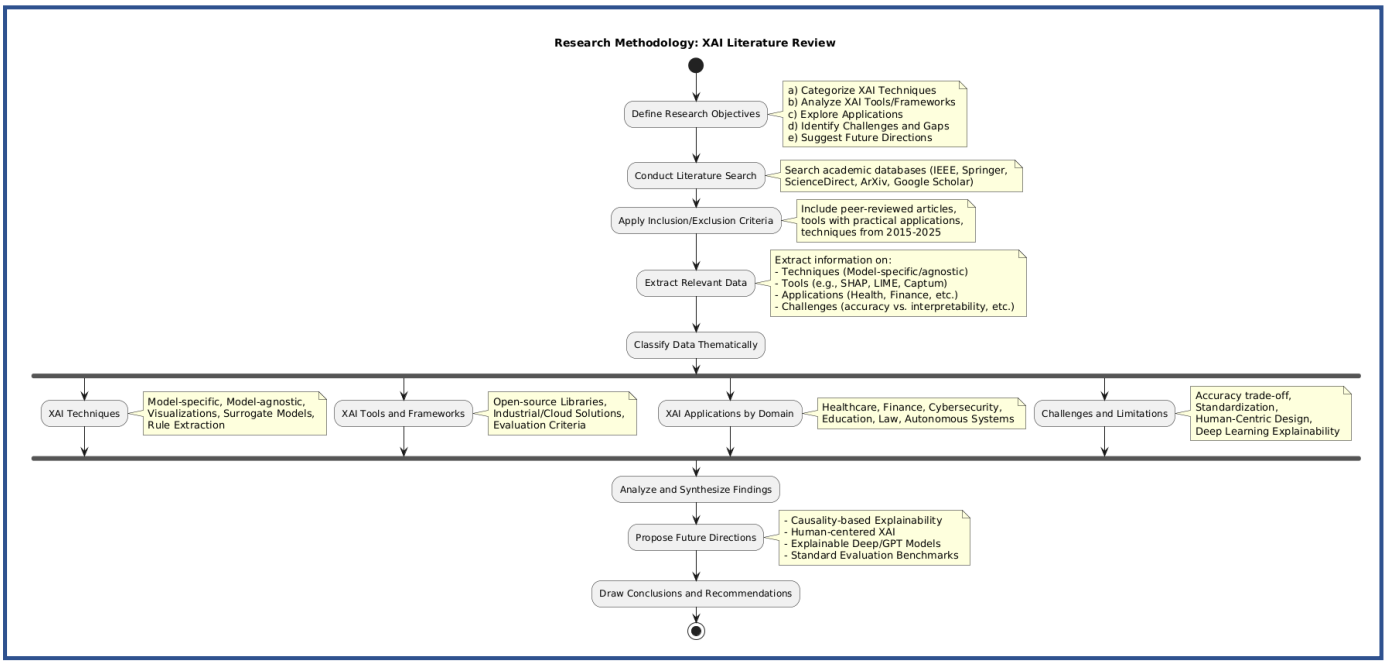

Figure 1: Research Methodology: XAI Literature Review

2 Fundamentals of Explainable Artificial Intelligence (XAI)

2.1 Definition and Concepts

Explainable Artificial Intelligence (XAI) refers to a set of methods and techniques designed to make the behaviour and decisions of AI systems understandable to humans [5]. In contrast to traditional black-box models, whose internal workings are largely inaccessible and non-transparent, XAI aims to provide human-comprehensible justifications for model outputs. These explanations help users answer critical questions such as: *Why did the model make a certain prediction? What input features influenced the outcome? How would a slight change in input alter the result?*

Explainability is especially essential when AI is used in high-stakes domains where decisions must be transparent, auditable, and fair [6]. It also plays a key role in facilitating trust, enhancing user adoption, supporting debugging and model improvement, and ensuring legal and regulatory compliance.

2.2 Importance of XAI

The growing demand for interpretability in AI models stems from several overlapping ethical, legal, and practical imperatives:

a) **Ethical Considerations:** AI systems can inadvertently reinforce bias, make discriminatory decisions, or operate in ways that are opaque and unjust. Explainability ensures that AI systems are accountable and respect human values [7].

b) **Legal and Regulatory Compliance:** Regulations such as the European Union's General Data Protection Regulation (GDPR) include provisions like the "right to explanation," requiring transparency in automated decision-making processes that affect individuals [8].

c) **Practical Implications:** For developers, explainable models enable easier debugging and optimization. For businesses and organizations, it supports informed decision-making and reduces the risks of reputational and operational damage due to AI errors [9].

In sum, XAI is a foundational requirement for responsible AI, ensuring that AI systems are not only intelligent but also understandable, fair, and trustworthy.

Tech-Sphere Journal of Pure and Applied Sciences (TSJPAS)
A Subsidiary of Tech-Sphere Multidisciplinary International Journal (TSMIJ)
Akinsiku. Vol 2, Issue 1, 2025 Publication Edition
ISSN: 3092-9598

2.3 Types of Explanations

Explanations in XAI can be categorized along two primary axes: Global vs Local, and Post-hoc vs Intrinsic.

Global vs Local Explanations

i. Global explanations offer a high-level understanding of the entire model's structure, behaviour, and logic. They aim to describe how the model makes decisions in general. For example, a decision tree may be completely interpretable as a global model [10].
ii. Local explanations focus on interpreting individual predictions. They provide insights into why a particular input led to a specific output. Tools like Local Interpretable Model-agnostic Explanations (LIME) and SHAP are commonly used for generating such explanations [11].

Post-hoc vs Intrinsic Explainability

i. Intrinsic explainability is built into the model itself. Models like linear regression, decision trees, and rule-based systems are inherently interpretable because their structures allow direct insight into how inputs relate to outputs [2].
ii. Post-hoc explanations are applied after model training, typically on complex models like neural networks or ensemble methods. These explanations are external approximations and include feature attribution, visualizations, counterfactuals, and surrogate models [12].

2.4 Stakeholders of Explanations

The design and delivery of explanations must consider the diverse needs of different stakeholders, each of whom interacts with AI systems from a unique perspective:

i. **Developers and Data Scientists:** Require detailed and technically grounded explanations to debug, refine, and optimize models. For them, interpretability is essential for model transparency and iterative improvement [13].
ii. **Domain Experts:** Professionals such as doctors, bankers, or engineers need domain-relevant explanations that support decision validation and alignment with real-world knowledge [14].
iii. **Regulators and Policymakers:** Require explanations that support auditability, compliance, and risk assessment, ensuring that models adhere to legal and ethical standards.
iv. **End Users and Consumers:** Often non-technical, they require simple, intuitive explanations that foster trust and clarity in how decisions affecting them were made like, loan approvals and job screening.

Understanding these stakeholder perspectives is key to designing effective XAI systems that are not only technically robust but also socially aligned.

3 XAI Techniques: A Thematic Review

Explainable Artificial Intelligence (XAI) employs a variety of methods to illuminate the inner workings of machine learning models. These techniques can be classified thematically into model-specific, model-agnostic, visualization-based, surrogate models, and rule extraction or feature importance approaches. Each category addresses different interpretability needs, catering to various model types, user expertise levels, and explanation purposes.

3.1 Model-Specific Techniques

Model-specific techniques are inherently interpretable and designed to provide built-in explainability. These models are often used in applications where transparency is prioritized over predictive performance.

a) **Decision Trees:** These are among the most transparent models, as they structure decisions through a series of if-then-else rules. Their branching paths clearly show how different input features affect the final prediction [15].

Tech-Sphere Journal of Pure and Applied Sciences (TSJPAS)
A Subsidiary of Tech-Sphere Multidisciplinary International Journal (TSMIJ)
Akinsiku. Vol 2, Issue 1, 2025 Publication Edition
ISSN: 3092-9598

b) **Generalized Additive Models (GAMs):** GAMs extend linear models by allowing non-linear relationships between the independent variables and the response variable, while maintaining interpretability. Each feature contributes additively and independently, making it easier to understand the effect of each input [16].

c) **Logistic Regression and Naïve Bayes:** Though simplistic, these models provide clear insights into how individual features contribute to the output probability, making them inherently interpretable [17].

These models are often preferred in domains like healthcare or risk assessment where regulatory compliance and human validation are critical

3.2 Model-Agnostic Techniques

Model-agnostic methods are designed to work with any predictive model, regardless of its complexity or architecture. They are typically applied after model training to explain either individual predictions or the global behaviour of the model.

a) **LIME (Local Interpretable Model-Agnostic Explanations):** LIME approximates the local behaviour of any black-box model by fitting an interpretable model (like a linear regression) around a specific prediction. It perturbs the input data and observes output changes to determine which features are most influential [18].

b) **SHAP (SHapley Additive exPlanations):** SHAP leverages principles from cooperative game theory (specifically Shapley values) to quantify the contribution of each feature to a model's output. It offers consistent and theoretically sound explanations, making it one of the most popular XAI techniques [19].

Both LIME and SHAP are widely adopted due to their flexibility, ease of use, and ability to explain complex models like ensemble trees and neural networks.

3.3 Visualization-Based Techniques

Visualization techniques play a vital role in interpreting AI models, particularly deep learning architectures, by translating complex model behaviours into forms that are understandable to humans [20]. Among these, saliency maps are widely used in computer vision tasks to highlight regions of an input image that significantly influenced the model's prediction, with methods like Gradient-weighted Class Activation Mapping (Grad-CAM) commonly applied to explain convolutional neural networks (CNNs) [21]. Partial Dependence Plots (PDPs) provide another perspective by illustrating the relationship between specific input features and the predicted outcome, averaging out the effects of other variables; this is particularly useful for interpreting models such as gradient-boosted decision trees. Feature attribution plots, including bar charts of SHAP values, offer clear insight into how much each feature contributes to individual predictions or the model's overall behaviour across the dataset [22]. These visualization tools are essential not only for exploratory analysis but also for fostering user trust, especially in domains where data is non-textual, such as image and audio processing, by making opaque model decisions more transparent and interpretable [23].

3.4 Surrogate Models

Surrogate models are simplified, interpretable models trained to approximate the behaviour of more complex and opaquer machine learning models, offering an accessible lens through which users can understand the decisions of black-box systems [24]. For example, a decision tree can be trained to replicate the predictions of a neural network or ensemble model on a given dataset, allowing analysts to examine the surrogate and gain insight into the logic of the original model without directly interacting with its complexity. These surrogate models can be applied globally to represent the overall behaviour of the original model, or locally to explain specific individual predictions. While surrogate models are valuable for generating intuitive explanations and enhancing model transparency, they must be interpreted with caution, as their approximations may not fully capture the nuances of the original model, particularly in regions of the input space characterized by sparse data or intricate decision boundaries [2].

Tech-Sphere Journal of Pure and Applied Sciences (TSJPAS)
A Subsidiary of Tech-Sphere Multidisciplinary International Journal (TSMIJ)
Akinsiku. Vol 2, Issue 1, 2025 Publication Edition
ISSN: 3092-9598

3.5 Rule Extraction and Feature Importance

This class of techniques aims to derive human-readable rules or quantify the influence of each input feature.

a) **Anchor Explanations:** Introduced by Ribeiro et al., anchors are high-precision rules that sufficiently "anchor" a prediction. If the anchor conditions hold true, the model is highly likely to produce the same output regardless of other features [25].
b) **Counterfactual Explanations:** These explanations answer "what if" questions. For instance: What minimal change to the input would have resulted in a different prediction? Counterfactuals are especially useful for decision-making in loan approval, hiring, and healthcare, where users may want actionable insights [26].
c) **Feature Importance Scores**: Many algorithms offer built-in mechanisms for ranking features (e.g., Gini importance in Random Forests, coefficients in linear models). These scores help identify which features are most influential to the model's decisions [27].

These techniques emphasize transparency, causality, and human actionability, and are frequently employed in fairness-aware and accountable AI systems.

Each XAI technique offers unique advantages depending on the context of application, model architecture, and target audience. In practice, a hybrid of these techniques is often used to achieve robust and user-aligned explainability

4 XAI Tools and Frameworks

The rapid advancement of Explainable Artificial Intelligence (XAI) has led to the development of various tools and frameworks that enable developers, researchers, and organizations to build and assess interpretable models. These tools fall into two broad categories: *open-source libraries*, which provide model-agnostic or model-specific interpretability methods [28], and *industrial/cloud-based platforms*, which embed XAI functionalities into enterprise-grade machine learning pipelines. This section reviews prominent tools and discusses their features, use cases, and evaluation criteria.

4.1 Open-source Libraries

A diverse set of robust open-source libraries has emerged to enhance transparency and interpretability in machine learning, offering both global and local explanations across various model types and seamlessly integrating with widely used ML frameworks. One of the most prominent is SHAP (SHapley Additive exPlanations), a Python library grounded in cooperative game theory that generates consistent and theoretically sound explanations for any model, with particular strength in tree-based algorithms such as XGBoost, LightGBM, and CatBoost, providing both global overviews and local attributions through intuitive visualizations [29]. LIME, on the other hand, explains individual predictions by fitting a simple, interpretable model (e.g., linear regression) locally around the prediction of interest, making it practical for a wide variety of classifiers and regressors [30]. InterpretML, developed by Microsoft, supports both inherently interpretable (glass-box) and complex black-box models by integrating explanation techniques like SHAP, LIME, and Generalized Additive Models (GAMs) into a cohesive, interactive visualization framework [31]. AIX360 (AI Explainability 360), from IBM, is a comprehensive Python toolkit that offers numerous post-hoc explanation algorithms along with metrics to evaluate the quality and utility of explanations for different user groups [32]. For PyTorch users, Captum provides gradient-based attribution methods such as Integrated Gradients, Saliency Maps, and DeepLIFT, making it particularly effective for interpreting deep neural networks in domains like computer vision and natural language processing [33]. Alibi, developed by Seldon, supports model-agnostic techniques including anchor explanations, counterfactuals, and contrastive methods, and is well-suited for production ML pipelines using TensorFlow or PyTorch [34]. Collectively, these open-source libraries empower researchers and practitioners to build transparent, reproducible, and customizable XAI workflows suitable for both academic inquiry and industrial deployment.

Tech-Sphere Journal of Pure and Applied Sciences (TSJPAS)
A Subsidiary of Tech-Sphere Multidisciplinary International Journal (TSMIJ)
Akinsiku. Vol 2, Issue 1, 2025 Publication Edition
ISSN: 3092-9598

4.2 Industrial and Cloud-based Solutions

Tech giants and cloud service providers have also introduced enterprise-ready platforms with built-in XAI capabilities. These tools are integrated into ML Ops workflows and support compliance, transparency, and auditability at scale.

a) **Google's What-If Tool**: A visual and interactive platform for TensorFlow models within TensorBoard. It enables users to test counterfactuals, analyse fairness, and assess model performance across subgroups, all without writing code [35].
b) **Microsoft Fairlearn:** Focused on both fairness and interpretability, Fairlearn integrates with scikit-learn and provides dashboards to assess disparities and visualize feature contributions. It is often used in regulated industries like banking and healthcare [36].
c) **AWS SageMaker Clarify**: A component of Amazon SageMaker that detects bias and provides explainability through SHAP-based feature attributions [37]. Clarify supports tabular, text, and image models, helping organizations meet explainability requirements in production environments.

These tools are especially valuable for organizations deploying AI at scale, providing end-to-end capabilities for model training, deployment, monitoring, and interpretation.

4.3 Comparison and Evaluation Criteria

Selecting the appropriate XAI tool requires evaluating how well it meets specific technical, ethical, and operational needs. The following criteria are commonly used to compare and assess XAI tools and frameworks:

a) **Fidelity:** Measures how accurately the explanation method reflects the true behaviour of the model. High-fidelity explanations are essential for gaining trustworthy insights.
b) **Complexity:** Refers to the cognitive load imposed by the explanation. Simpler explanations are more accessible to non-technical users but may sacrifice detail.
c) **Stability:** Indicates whether small changes in input lead to consistent explanations. Instability can reduce user trust and reliability of the explanation method.
d) **Generalizability:** Describes the applicability of the tool across different types of models, tasks, and datasets. Highly generalizable tools are preferred in environments where diverse models are used.

Other considerations include computational cost, integration ease, visualization support, and regulatory alignment. No single tool fits all use cases; the choice should be guided by the target audience, domain of application, and deployment constraints.

Here is a flowchart that visually organizes the XAI Tools and Frameworks section as depicted on Figure 2. It categorizes the tools into:

a) Open-source Libraries (e.g., SHAP, LIME, Captum)
b) Industrial/Cloud-based Solutions (e.g., Google's What-If Tool, AWS Clarify)
c) Evaluation Criteria (e.g., Fidelity, Complexity, Stability, Generalizability)

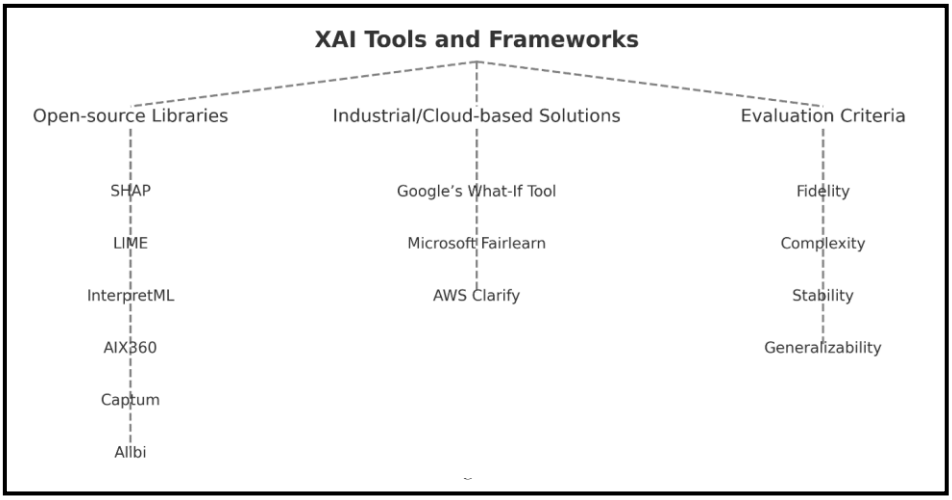

Figure 2: XAI Tools and Frameworks

In conclusion, the landscape of XAI tools is rich and continually evolving. A clear understanding of each tool's strengths and limitations is essential for designing AI systems that are not only powerful but also transparent, fair, and responsible.

5 Applications of XAI in Various Domains

The adoption of Artificial Intelligence (AI) across diverse industries has amplified the need for transparent and trustworthy decision-making. Explainable Artificial Intelligence (XAI) plays a pivotal role in enabling human users to understand, trust, and effectively manage AI systems. This section highlights the practical application of XAI across key sectors, each with unique interpretability requirements and high stakes.

5.1 Healthcare

In healthcare, where decisions can be life-altering, explainability is not just desirable but essential. AI models are increasingly used for diagnostic imaging, disease prediction, treatment planning, and patient risk assessment. However, the black-box nature of many models limits their acceptance among clinicians [38].

XAI enables:

i. Model transparency in medical diagnosis, such as highlighting specific regions in an MRI scan that contributed to a prediction of a brain tumour.
ii. Prognostic analysis by explaining risk scores in chronic diseases (e.g., cardiovascular risk calculators).
iii. Clinical trust and validation, allowing practitioners to verify whether AI decisions align with established medical knowledge.

Tools like SHAP and LIME are often employed to interpret deep learning models in radiology and genomics, offering both local and global insights into the AI's decision process.

5.2 Finance

Financial services demand fairness, accountability, and regulatory compliance. XAI is used to interpret AI models employed for:

i. Credit scoring: Explaining why a loan application was approved or denied.
ii. Fraud detection: Understanding why a particular transaction was flagged as suspicious.
iii. Algorithmic trading: Providing insight into trading strategy decisions generated by AI.

With increasing legal oversight like, GDPR, Basel regulations, XAI helps financial institutions maintain transparency and avoid biases that may inadvertently discriminate against individuals based on protected attributes such as gender or race. SHAP values can identify which features like credit history, income level influenced a loan decision, allowing both auditors and customers to understand the rationale behind outcomes [39].

5.3 Autonomous Systems

Autonomous systems, such as self-driving vehicles and delivery drones, rely on complex AI models to perceive their environment, plan routes, and make real-time decisions [40]. A single error can result in accidents or legal consequences. XAI supports:

a) Interpretation of perception modules (e.g., object recognition systems explaining why an obstacle was classified as a pedestrian).
b) Transparency in decision-making (e.g., explaining why a car chose to brake or swerve in a specific situation).
c) Failure analysis and debugging during system testing and post-incident reviews.

Techniques like saliency maps and scenario simulations help engineers understand how sensor inputs (camera, lidar, radar) contribute to control decisions.

5.4 Cybersecurity

AI-driven systems in cybersecurity are used to detect anomalies, intrusions, and malware [41]. However, their complexity often obscures the reasoning behind alerts, leading to "alert fatigue" and distrust from security analysts.
With XAI:

a) Threat detection systems become interpretable, allowing analysts to understand what triggered an alert (e.g., a spike in port scanning or suspicious login attempts).
b) False positives can be reduced, as explainability helps refine rule-based tuning and improves detection logic.
c) Real-time decision support becomes actionable, improving incident response and forensic investigation.

XAI tools like LIME, Alibi, and tree-based explanations are commonly integrated into Security Information and Event Management (SIEM) systems.

5.5 Education and EdTech

In education, AI models are applied for adaptive learning, performance analytics, and intelligent tutoring. XAI enhances educational AI systems by:

a) Personalizing learning paths and providing understandable feedback to students on why certain recommendations (e.g., exercises or resources) were made.
b) Supporting educators by explaining student engagement and performance predictions.
c) Improving model accountability, especially when used for grading or admissions decisions.

For example, a student's low score prediction can be explained by factors such as poor quiz performance or low engagement time, prompting personalized intervention [42].

5.6 Law and Policy

In legal and policy domains, the use of AI for risk assessments, sentencing decisions, and resource allocation demands maximum transparency and fairness. Black-box decisions in this context can lead to significant ethical and legal implications [43].
XAI addresses:

a) Legal accountability, ensuring that AI-based decisions can be explained and defended in court or legal audits.
b) Regulatory compliance, by fulfilling requirements for transparency in automated systems (e.g., the right to explanation under GDPR).
c) Bias detection and mitigation, preventing systemic discrimination in tools used for parole decisions or public service allocation.

Rule-based models and counterfactual explanations are particularly valued in this domain due to their alignment with legal reasoning and precedent analysis.
In summary, XAI is not a one-size-fits-all solution. Each domain applies XAI differently based on contextual needs, stakeholder expertise, and the nature of the decisions involved. As AI becomes more pervasive, XAI will continue to serve as a foundational element in ensuring that AI systems are understandable, fair, and human-centric.

6 Challenges and Limitations in XAI

Despite significant advances, Explainable Artificial Intelligence (XAI) faces a number of persistent challenges and limitations that hinder its broader adoption, effectiveness, and standardization across disciplines. These challenges span technical, human, and contextual dimensions, and highlight the complexity of achieving meaningful transparency in AI systems. This section explores some of the most pressing concerns [44].

Tech-Sphere Journal of Pure and Applied Sciences (TSJPAS)
A Subsidiary of Tech-Sphere Multidisciplinary International Journal (TSMIJ)
Akinsiku. Vol 2, Issue 1, 2025 Publication Edition
ISSN: 3092-9598

6.1 Trade-off Between Accuracy and Interpretability

One of the most fundamental challenges in XAI is the inherent trade-off between model complexity (accuracy) and interpretability. High-performing models, such as deep neural networks, ensemble trees, and large language models, often achieve superior accuracy but at the cost of being opaque or non-intuitive [45]. Conversely, simpler models like decision trees and linear regression are more interpretable but may lack the expressive power to capture intricate patterns in data. This trade-off forces practitioners to choose between performance and transparency, particularly in high-stakes applications like healthcare and criminal justice where both are critical. Recent efforts in hybrid modelling and self-explaining models aim to balance this tension, but the problem remains largely unresolved.

6.2 Standardization and Benchmarking

The XAI landscape is currently fragmented, with a multitude of techniques, metrics, and tools emerging from diverse research communities. There is no universally accepted standard for what constitutes a "good" explanation, nor agreed-upon benchmarks for evaluating different methods.

Challenges include:

a) Lack of unified evaluation metrics (e.g., how to objectively measure faithfulness, robustness, or usability of explanations).
b) Tool interoperability issues, especially when integrating different XAI methods into production pipelines.
c) Reproducibility concerns, as explanations can vary across models, inputs, or even random seeds.

Without standardized frameworks, it becomes difficult to compare approaches, replicate findings, or assess regulatory compliance across domains.

6.3 Human Factors and Cognitive Load

XAI is ultimately meant for human users—but not all explanations are equally useful or understandable to all audiences. A technically accurate explanation may still fail if it is not cognitively accessible or aligned with a user's domain expertise [46].

Key issues include:

a) Cognitive overload from overly detailed or technical explanations.
b) Misinterpretation of visualizations or feature attributions, especially by non-technical stakeholders.
c) Trust calibration, where poor or inconsistent explanations may reduce user trust or, conversely, lead to over-trust in unreliable systems.

Human-centred design, usability studies, and user-specific tailoring of explanations are essential, but currently underdeveloped areas in XAI research and practice

6.4 Domain-Specific Constraints

XAI solutions are not universally applicable across all fields. Different application domains impose unique constraints on what explanations are acceptable, meaningful, or actionable.

For instance:

i. In healthcare, explanations must align with medical reasoning and terminology, often requiring strict validation by clinical experts.
ii. In finance and law, explanations must be auditable, rule-based, and legally defensible.
iii. In cybersecurity, explanations must support real-time decision-making under pressure, often with low tolerance for false positives.

Such domain-specific requirements necessitate customization of XAI techniques, complicating the development of general-purpose explainability frameworks.

Tech-Sphere Journal of Pure and Applied Sciences (TSJPAS)
A Subsidiary of Tech-Sphere Multidisciplinary International Journal (TSMIJ)
Akinsiku. Vol 2, Issue 1, 2025 Publication Edition
ISSN: 3092-9598

6.5 Explainability in Deep Learning and Generative Models

Modern AI systems increasingly rely on deep learning architectures and generative models like GPT, DALL·E and diffusion models, which are notoriously difficult to interpret [47]. These models consist of millions (or billions) of parameters and learn abstract representations that do not map directly to human-intuitive features.

Challenges include:

a) Layer-wise opacity, where decisions are influenced by internal representations that defy simple attribution.
b) Emergent behaviours in large models, where complex reasoning appears without being explicitly programmed.
c) Lack of causality, as explanations in deep models often describe correlations rather than underlying causes.
d) Interpretability degradation when combining multiple models (e.g., in pipelines or multi-modal systems).

While tools like saliency maps, attention mechanisms, and feature attribution methods have made some progress, truly interpretable deep learning remains an open research problem.

In conclusion, while XAI offers promising paths toward responsible and transparent AI, it is not without significant limitations. Addressing these challenges requires interdisciplinary collaboration, standardization efforts, and human-centric approaches that balance technical fidelity with usability and ethical responsibility.

7 Conclusion

7.1 Summary of Key Insights

This literature review has provided a comprehensive exploration of Explainable Artificial Intelligence (XAI), emphasizing its growing importance in making AI systems transparent, accountable, and trustworthy. The review began by introducing the motivation for XAI, stemming from the black-box nature of many high-performing models, and outlined the ethical, legal, and technical imperatives for interpretability. A thematic overview of XAI techniques highlighted both model-specific approaches (e.g., decision trees, GAMs) and model-agnostic methods (e.g., SHAP, LIME), along with visualization strategies and surrogate modelling. These methods offer varying degrees of transparency, fidelity, and usability across different use cases. We examined a diverse set of tools and frameworks, including open-source libraries such as Captum and AIX360, and enterprise solutions like AWS Clarify and Google's What-If Tool. Evaluation criteria, such as fidelity, stability, complexity, and generalizability, were discussed as key benchmarks for selecting and deploying these tools effectively. The real-world applications of XAI span across critical domains such as healthcare, finance, autonomous systems, cybersecurity, education, and law, each presenting unique interpretability challenges and requirements. Additionally, the paper outlined significant challenges and limitations in current XAI approaches, including the accuracy–interpretability trade-off, lack of standardization, and cognitive load on users. Lastly, we looked toward the future of XAI, identifying promising research directions such as the integration with causal inference, deployment in federated learning and edge AI, development of interactive human-centred systems, and advancing explainability in complex models like reinforcement learners and large language models.

7.2 Significance of XAI for Responsible AI

Explainable AI is not merely a technical convenience, it is a cornerstone of responsible AI. As AI systems increasingly influence decisions that affect human lives and society, the ability to explain, justify, and audit those decisions becomes paramount. XAI supports:

a) Trust and adoption, by helping users understand and rely on AI systems.
b) Fairness and accountability, by exposing biases and enabling compliance with ethical and legal standards.
c) Safety and performance, by facilitating debugging, model validation, and user feedback loops.

In this context, explainability is central to the ethical development and deployment of AI systems that are aligned with human values and social expectations

7.3 Final Thoughts on the Path Ahead

The journey toward universally explainable AI is still unfolding. While significant strides have been made, many challenges remain, especially in standardizing evaluation, scaling interpretability to modern architectures, and tailoring explanations to human needs. To truly unlock the potential of AI in critical sectors, future XAI research must be interdisciplinary, context-aware, and user-centric. Collaboration between computer scientists, domain experts, ethicists, policymakers, and designers will be essential in shaping the next generation of AI systems that are not only powerful but also transparent, equitable, and accountable. In conclusion, the path forward for XAI is one of convergence, merging technical sophistication with human understanding, to ensure that artificial intelligence remains an empowering force for all.

References

[1] K. Razzaq and M. Shah, "Machine learning and deep learning paradigms: From techniques to practical applications and research frontiers," *Computers*, vol. 14, no. 3, p. 93, 2025.
[2] V. Hassija *et al.*, "Interpreting black-box models: a review on explainable artificial intelligence," *Cognit. Comput.*, vol. 16, no. 1, pp. 45–74, 2024.
[3] E. Utomi, A. S. Osifowokan, A. A. Donkor, and I. A. Yowetu, "Evaluating the Impact of Data Protection Compliance on AI Development and Deployment in the US Health sector," *World J. Adv. Res. Rev.*, vol. 24, no. 2, pp. 1100–1110, 2024.
[4] N. Rane, S. P. Choudhary, and J. Rane, "Acceptance of artificial intelligence: key factors, challenges, and implementation strategies," *J. Appl. Artif. Intell.*, vol. 5, no. 2, pp. 50–70, 2024.
[5] V. Chamola, V. Hassija, A. R. Sulthana, D. Ghosh, D. Dhingra, and B. Sikdar, "A review of trustworthy and explainable artificial intelligence (XAI)," *IEEe Access*, vol. 11, pp. 78994–79015, 2023.
[6] R. David, H. Shankar, P. Kura, K. Kowtarapu, and S. Karkuzhali, "Advancement in Explainable AI: Bringing Transparency and Interpretability to Machine Learning Models for Use in High-Stakes Decisions," in *2025 International Conference on Emerging Smart Computing and Informatics (ESCI)*, IEEE, 2025, pp. 1–6.
[7] L. Emma, "The Ethical Implications of Artificial Intelligence: A Deep Dive into Bias, Fairness, and Transparency," *Retrieved from Emma, L.(2024). Ethical Implic. Artif. Intell. A Deep Dive into Bias, Fairness, Transpar.*, 2024.
[8] A. Nandan Prasad, "Regulatory Compliance and Risk Management," in *Introduction to Data Governance for Machine Learning Systems: Fundamental Principles, Critical Practices, and Future Trends*, Springer, 2024, pp. 485–624.
[9] A. Ayele, K. K. Pantangi, and S. A. Olugbabi, "Developer Perspectives on AI-Driven Code Debugging In Financial Systems," 2025.
[10] R. Abi, "Ethical and Explainable AI in Data Science for Transparent Decision-Making Across Critical Business Operations".
[11] S. Ahmed, M. S. Kaiser, M. S. Hossain, and K. Andersson, "A comparative analysis of LIME and SHAP interpreters with explainable ML-based diabetes predictions," *IEEE access*, vol. 13, pp. 37370–37388, 2024.
[12] A. Madsen, S. Reddy, and S. Chandar, "Post-hoc interpretability for neural nlp: A survey," *ACM Comput. Surv.*, vol. 55, no. 8, pp. 1–42, 2022.
[13] K. Sankaran, "Data science principles for interpretable and explainable AI," *arXiv Prepr. arXiv2405.10552*, 2024.
[14] I. A. Zahid *et al.*, "Explainability, Robustness, and Fairness in User-Centric Intelligent Systems: A Systematic Review," *IEEE Trans. Emerg. Top. Comput. Intell.*, 2025.
[15] K. Kanamori, T. Takagi, K. Kobayashi, and Y. Ike, "Counterfactual explanation trees: Transparent and consistent actionable recourse with decision trees," in *International Conference on Artificial Intelligence and Statistics*, PMLR, 2022, pp. 1846–1870.
[16] P. Zschech, S. Weinzierl, N. Hambauer, S. Zilker, and M. Kraus, "GAM (e) changer or not? An evaluation of interpretable machine learning models based on additive model constraints," *arXiv Prepr. arXiv2204.09123*, 2022.
[17] A. Tursunalieva, D. L. J. Alexander, R. Dunne, J. Li, L. Riera, and Y. Zhao, "Making sense of machine learning: a review of interpretation techniques and their applications," *Appl. Sci.*, vol. 14, no. 2, p. 496, 2024.
[18] M. R. Zafar and N. Khan, "Deterministic local interpretable model-agnostic explanations for stable explainability," *Mach. Learn. Knowl. Extr.*, vol. 3, no. 3, pp. 525–541, 2021.
[19] A. Heuillet, F. Couthouis, and N. Díaz-Rodríguez, "Collective explainable AI: Explaining cooperative strategies and agent contribution in multiagent reinforcement learning with shapley values," *IEEE Comput. Intell. Mag.*, vol. 17, no. 1, pp. 59–71, 2022.
[20] A. K. Reddy, S. K. Thota, V. Saini, S. Chitta, and S. G. R. Bojja, "Bridging AI and Human Understanding: Interpretable Deep Learning in Practice," *J. Informatics Educ. Res.*, vol. 4, p. 3706, 2024.
[21] T. Gomez and H. Mouchère, "Computing and evaluating saliency maps for image classification: a tutorial," *J. Electron. Imaging*, vol. 32, no. 2, p. 20801, 2023.
[22] C. Molnar *et al.*, "Relating the partial dependence plot and permutation feature importance to the data generating process," in *World Conference on Explainable Artificial Intelligence*, Springer, 2023, pp. 456–479.
[23] D. Herrera-Poyatos *et al.*, "An overview of model uncertainty and variability in LLM-based sentiment analysis. Challenges, mitigation strategies and the role of explainability," *arXiv Prepr. arXiv2504.04462*, 2025.
[24] E. ŞAHiN, N. N. Arslan, and D. Özdemir, "Unlocking the black box: an in-depth review on interpretability, explainability, and reliability in deep learning," *Neural Comput. Appl.*, vol. 37, no. 2, pp. 859–965, 2025.
[25] Y. Wang, "A comparative analysis of model agnostic techniques for explainable artificial intelligence," *Res. Reports Comput. Sci.*, pp. 25–33, 2024.
[26] L. De Schutter and D. De Cremer, "How counterfactual fairness modelling in algorithms can promote ethical decision-making," *Int. J. Human–Computer Interact.*, vol. 40, no. 1, pp. 33–44, 2024.
[27] R. Dunne *et al.*, "Thresholding Gini variable importance with a single-trained random forest: An empirical Bayes approach," *Comput. Struct. Biotechnol. J.*, vol. 21, pp. 4354–4360, 2023.
[28] L. Galera Alfaro, "Generating an Interpretable Ranking Model: Exploring the Power of Local Model-Agnostic Interpretability for Ranking Analysis." 2023.
[29] M. Li, H. Sun, Y. Huang, and H. Chen, "Shapley value: from cooperative game to explainable artificial intelligence," *Auton. Intell. Syst.*, vol. 4, no. 1, p. 2, 2024.

[30] M. Henninger and C. Strobl, "Interpreting machine learning predictions with LIME and Shapley values: theoretical insights, challenges, and meaningful interpretations," *Behaviormetrika*, vol. 52, no. 1, pp. 45–75, 2025.
[31] R. Dwivedi *et al.*, "Explainable AI (XAI): Core ideas, techniques, and solutions," *ACM Comput. Surv.*, vol. 55, no. 9, pp. 1–33, 2023.
[32] V. Arya *et al.*, "AI explainability 360: Impact and design," in *Proceedings of the AAAI conference on artificial intelligence*, 2022, pp. 12651–12657.
[33] D. Jin, E. Sergeeva, W. Weng, G. Chauhan, and P. Szolovits, "Explainable deep learning in healthcare: A methodological survey from an attribution view," *WIREs Mech. Dis.*, vol. 14, no. 3, p. e1548, 2022.
[34] G. Meller, "Explainable Artificial Intelligence: A Study of Methods, Applications, and Future Directions".
[35] P. Mishra, "AI model fairness using a what-if scenario," in *Practical Explainable AI Using Python: Artificial Intelligence Model Explanations Using Python-based Libraries, Extensions, and Frameworks*, Springer, 2021, pp. 229–242.
[36] B. Johnson and Y. Brun, "Fairkit-learn: a fairness evaluation and comparison toolkit," in *Proceedings of the ACM/IEEE 44th international conference on software engineering: Companion proceedings*, 2022, pp. 70–74.
[37] M. Argymbay, S. Khan, N. Ahmad, M. Salih, and Y. Mamatjan, "A Smart Recommender System for Stroke Risk Assessment with an Integrated Strokebot," *J. Med. Biol. Eng.*, vol. 44, no. 6, pp. 799–808, 2024.
[38] A. R. R. Salammagari and G. Srivastava, "Artificial intelligence inhealthcare: Revolutionizing disease diagnosis and treatmentplanning," *Int. J. Res. Comput. Appl. Inf. Technol*, vol. 7, pp. 41–53, 2024.
[39] N. Rane, S. Choudhary, and J. Rane, "Explainable Artificial Intelligence (XAI) approaches for transparency and accountability in financial decision-making," *Available SSRN 4640316*, 2023.
[40] G. Bathla *et al.*, "Autonomous vehicles and intelligent automation: Applications, challenges, and opportunities," *Mob. Inf. Syst.*, vol. 2022, no. 1, p. 7632892, 2022.
[41] M. I. Khan, A. Arif, and A. R. A. Khan, "AI-Driven Threat Detection: A Brief Overview of AI Techniques in Cybersecurity," *BIN Bull. Informatics*, vol. 2, no. 2, pp. 248–261, 2024.
[42] C. Conati, O. Barral, V. Putnam, and L. Rieger, "Toward personalized XAI: A case study in intelligent tutoring systems," *Artif. Intell.*, vol. 298, p. 103503, 2021.
[43] G. Chaudhary, "Unveiling the black box: Bringing algorithmic transparency to AI," *Masaryk Univ. J. Law Technol.*, vol. 18, no. 1, pp. 93–122, 2024.
[44] D. Minh, H. X. Wang, Y. F. Li, and T. N. Nguyen, "Explainable artificial intelligence: a comprehensive review," *Artif. Intell. Rev.*, vol. 55, no. 5, pp. 3503–3568, 2022.
[45] A. Assis, J. Dantas, and E. Andrade, "The performance-interpretability trade-off: A comparative study of machine learning models," *J. Reliab. Intell. Environ.*, vol. 11, no. 1, p. 1, 2025.
[46] B. Severes, C. Carreira, A. B. Vieira, E. Gomes, J. T. Aparício, and I. Pereira, "The human side of xai: Bridging the gap between ai and non-expert audiences," in *Proceedings of the 41st ACM International Conference on Design of Communication*, 2023, pp. 126–132.
[47] S. Bengesi, H. El-Sayed, M. K. Sarker, Y. Houkpati, J. Irungu, and T. Oladunni, "Advancements in generative AI: A comprehensive review of GANs, GPT, autoencoders, diffusion model, and transformers," *IEEe Access*, vol. 12, pp. 69812–69837, 2024.