

Review Article

Explainable Artificial Intelligence: A Comprehensive Review of Techniques, Applications, and Emerging Trends

Munyao Muia^{1*}, Jackson Kamiri²

¹Dept. of Computer Science, Murang'a University of Technology, Murang'a, Kenya

²Dept. of Information Technology, Murang'a University of Technology, Murang'a, Kenya

Corresponding Author: 

Received: 22/Jun/2025; Accepted: 24/Jul/2025; Published: 31/Aug/2025. | DOI: <https://doi.org/10.26438/ijsrcse.v13i4.740>

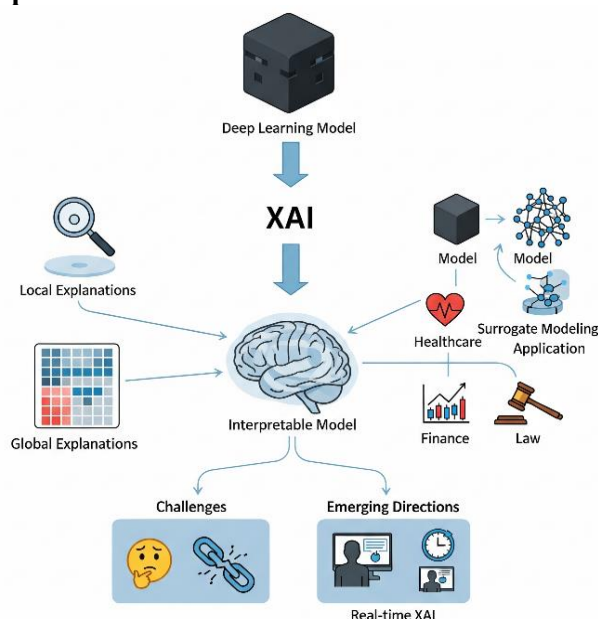


Copyright © 2025 by author(s). This is an Open Access article distributed under the terms of the [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/) which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited & its authors credited.

Abstract—The growing reliance on opaque deep learning models in critical domains has intensified the demand for transparency, accountability, and interpretability of artificial intelligence systems. Explainable Artificial Intelligence (XAI) seeks to address this through methods that clarify how models make predictions, fostering informed oversight and trust. This review synthesizes foundational concepts in XAI, distinguishing interpretability from explainability and contrasting intrinsic with post-hoc methods, as well as local with global explanations. It surveys major techniques such as feature attribution, gradient-based methods, surrogate models, counterfactual reasoning, and inherently interpretable architectures alongside widely used tools, frameworks, and evaluation metrics, including fidelity, human interpretability, robustness, and efficiency. Applications in healthcare, finance, law, and autonomous systems are discussed, highlighting domain-specific interpretability needs. Persistent challenges, including inconsistent terminology, subjective evaluation, and adversarial vulnerabilities, are examined. The review concludes with emerging directions in human-centered, real-time, and multimodal explanations, aiming to guide the design of XAI systems that are both technically sound and aligned with societal expectations.

Keywords— Explainable AI (XAI), Interpretability, Transparency in AI, Post-hoc Explanation, Human-Centered AI, Evaluation of XAI

Graphical Abstract



The graphical abstract visually summarizes this review on Explainable Artificial Intelligence (XAI), depicting how XAI enhances transparency and trust in opaque deep learning models. It highlights core techniques, their applications in domains such as healthcare, finance, law, and autonomous systems, key challenges of evaluation, robustness, and generalization, and future directions emphasizing human-centered design, regulatory alignment, and real-time explainability.

1. Introduction

Recent progress in machine learning, especially deep learning, has driven major innovations across diverse fields, including medicine, financial services, transportation, and legal systems. However, the increasing reliance on complex, opaque models has raised profound concerns about trust, accountability, and transparency in artificial intelligence (AI) [1]. These so-called "black-box" models, while highly accurate, often operate in ways that are incomprehensible to end-users and decision-makers [1]. In high-stakes

environments where decisions can impact human lives, livelihoods, or legal standing, the inability to understand how an AI system arrives at its conclusions poses serious ethical, legal, and operational risks.

1.1 Problem Statement

The increasing conflict between model accuracy and transparency has spurred the emergence of Explainable Artificial Intelligence (XAI), encompassing methods and frameworks designed to clarify how AI systems operate and make decisions, while striving to maintain their effectiveness [2]. As the demand for transparent and trustworthy AI intensifies, driven by regulatory frameworks, societal expectations, and the need for responsible AI adoption, XAI has evolved into a critical research frontier [3]. It addresses not only the technical challenges of interpreting complex models but also the broader goal of aligning AI behavior with human values and expectations.

Although considerable progress has been made in developing interpretability methods, the XAI field remains fragmented. Existing literature often focuses on narrow subsets of techniques, targets specific domains, or lacks rigorous comparative assessments. Moreover, existing reviews emphasize taxonomies or technical formulations without addressing the practical inferences of deploying these methods in real-world settings [4-6]. Therefore, there is a pressing need to bridge the gap between predictive accuracy and interpretability to foster responsible AI adoption, along with societal expectations for fairness and accountability in making explainability an essential component of trustworthy AI systems.

1.2 Objective of the Study

The primary objective of the review is to provide an in-depth and structured overview of the XAI landscape, beginning with conceptual foundations such as the distinction between interpretability and explainability, classifications based on scope and purpose, and the overarching goals of the field. The paper aims to review key XAI techniques, examining their theoretical underpinnings, practical applications, and trade-offs. Widely adopted software tools and frameworks that support scalable deployment will be analysed. Additionally, the paper will explore evaluation approaches, ranging from fidelity and stability to human-centered assessment strategies, and discuss benchmark datasets used for empirical validation.

Recognizing the importance of context, the paper aims to cover domain-specific applications of XAI in healthcare, finance, legal systems, and autonomous technologies, each posing unique interpretability challenges. The aim is also to examine major challenges such as balancing accuracy with interpretability, the absence of unified evaluation standards, and the susceptibility of explanations to adversarial attacks. Finally, the review will outline emerging trends such as human-in-the-loop systems, explainability for generative models, and the evolving role of policy and regulation. This study is significant in that it aims to serve as a comprehensive and critical resource for AI researchers, practitioners, and policy-makers aiming to develop, deploy, or regulate

explainable AI systems by integrating methodological depth with practical insights.

1.3 Organization of the Paper

This article is organized into the following sections which are as follows; Section 1 contains the introduction of the study, Section 2 contain the related work of the study, Section 3 contains a detailed review of explanation techniques, Section 4 covers prominent software tools and frameworks for XAI, Section 5 discusses evaluation metrics and benchmarks, Section 6 explores domain-specific applications, Section 7 presents the synthesized results and discussion and Section 8 concludes research work with future directions.

2. Related Work

Several studies have explored various aspects of explainable artificial intelligence (XAI), focusing on its conceptual foundations, methodologies, and applications. Early works established fundamental distinctions such as interpretability versus explainability, while others examined intrinsic versus post-hoc approaches and the trade-offs between accuracy and transparency. General-purpose surveys have provided taxonomies of XAI methods, whereas domain-specific reviews have emphasized medical, financial, and legal applications. Building on these efforts, this review consolidates existing knowledge while offering a broader synthesis that integrates technical foundations, evaluation methods, and emerging research trends across multiple domains.

At the heart of XAI lies a conceptual tension between two closely related ideas: interpretability and explainability. While frequently used interchangeably, these terms differ subtly but significantly. Interpretability generally describes how easily a person can comprehend the internal workings and logic of a model. [7]. Decision trees and linear regression models are inherently interpretable because their structure and functioning are straightforward and transparent. Explainability, on the other hand, relates to the extent to which external tools or techniques can render a model's decisions understandable. In practice, explainability often involves post-hoc methods that attempt to make sense of already-trained, opaque models [7]. Studies show that these methods do not necessarily open the model's internals but instead generate outputs, such as feature importance scores, visual heatmaps, or rule approximations, that humans can interpret.

Another fundamental dimension of XAI concerns the distinction between intrinsic and post-hoc explainability. Intrinsically interpretable models, such as rule-based systems or generalized additive models, are designed with transparency as a core attribute [8]. Their simplicity allows for direct inspection and understanding. However, this simplicity often comes at the cost of predictive performance, particularly in tasks involving high-dimensional or unstructured data. Post-hoc explainability, in contrast, involves applying explanation methods after the training of a black-box model [9]. These methods generate interpretations

without modifying the model itself, allowing practitioners to retain accuracy while gaining interpretive insights [8]. Notable examples are LIME (Local Interpretable Model-agnostic Explanations), SHAP (SHapley Additive exPlanations), and saliency maps, all of which provide varying degrees of understanding into how models generate their outputs.

The scope of an explanation, whether local or global, is another key axis along which XAI methods can be classified. Local explanations concentrate on specific predictions, clarifying the reasoning behind the model's decision for a particular input [10]. Techniques such as LIME and SHAP are widely used for this purpose, as they generate instance-specific feature attributions. Global explanations, on the contrary, describe the model's behavior across the entire input space [10]. These include approaches such as feature importance plots, global surrogate models, or aggregated attention weights in neural networks. While local explanations are helpful for end-user transparency and debugging, global explanations are essential for auditing model behavior, understanding structural biases, and gaining system-level insights.

Another way to categorize XAI approaches is by dividing them into model-specific and model-agnostic techniques. Model-specific approaches are tailored for certain types of models and leverage their internal mechanisms. [11]. For example, Grad-CAM is tailored for convolutional neural networks and leverages gradient information to generate visual explanations. In contrast, model-agnostic methods operate independently of the underlying model architecture. These include perturbation-based techniques such as LIME or sampling-based approaches like SHAP, which can be applied to any model type, such as random forests, neural networks, and support vector machines (SVMs) [11]. Model-agnostic techniques are valued for their flexibility and ease of integration, particularly in heterogeneous production environments where models vary in type and complexity.

The overarching goals of XAI extend beyond technical transparency. They include fostering user trust by enabling stakeholders to understand and verify model outputs; supporting model debugging by revealing failure modes, biases, or inconsistencies; guaranteeing amenability with various legal and ethical standards that demand explainability in automated decision-making; and enabling effective human-AI collaboration where users can engage with, contest, or override algorithmic decisions [2]. These goals make XAI a socio-technical effort that blends machine learning with domains such as human-computer interaction, cognitive science, and ethics. As such, developing meaningful explanations requires not only algorithmic consistency but also a deep understanding of user needs, domain constraints, and interpretive context.

These foundational studies collectively provide the basis for understanding XAI. However, they often address isolated aspects of the field. The present review extends this body of work by offering an integrated analysis of techniques, tools,

evaluation frameworks, domain-specific challenges, and future directions.

3. XAI Techniques

The development of explainability techniques has been driven by the need to ensure that deep learning models are more transparent and dependable, particularly in critical domains where decision justification is essential. These techniques vary widely in their theoretical motivations, operational mechanisms, and applicability across model types and data modalities. This section presents a detailed review of representative XAI techniques, organized by their methodological class. Each category is discussed in terms of its core principles, practical utility, and associated limitations.

3.1 Feature Attribution Techniques

Feature attribution methods seek to measure how much each input feature influences a model's output. A notable example is LIME (Local Interpretable Model-Agnostic Explanations), which explains individual predictions by approximating the complex model's behavior near a specific instance using a simpler, interpretable surrogate such as a linear model or a decision tree [12]. LIME provides interpretable local explanations by altering the input slightly and examining how these changes affect the output. Nonetheless, its dependence on random perturbations and surrogate models may lead to variability or inconsistency in the explanations, particularly in feature spaces that are in high dimensions [12].

SHAP (SHapley Additive exPlanations) is based on cooperative game theory and assigns each feature a value representing its marginal impact on a prediction [13]. It upholds key properties like consistency, local accuracy, and additivity, making it one of the most robust and theoretically well-founded attribution techniques. While SHAP offers robust and consistent explanations, its computational cost can be prohibitive, particularly for models with many input features or complex architectures [13]. TreeSHAP and other optimized implementations have been developed to improve scalability for specific model families, such as gradient boosting machines.

DeepLIFT (Deep Learning Important FeaTures) is another attribution technique specifically designed for deep neural networks. It evaluates each neuron's activation against a reference point and determines contribution scores guided by the deviations observed. This technique avoids some of the saturation problems observed in gradient-based methods and provides high-resolution explanations for layered networks [14]. Nonetheless, its interpretive quality is sensitive to the selected baseline, which is a recurring problem in attribution methods.

3.2 Gradient-Based Techniques

Gradient-based methods use the gradients of the model's output relative to its input to identify the features or regions that most impact a prediction. Saliency maps, among the earliest techniques in this class, visualize the magnitude of gradients to highlight input features that strongly affect the

output [15]. While easy to implement, saliency maps can be noisy and lack clarity, especially when gradients are sparse or highly non-linear.

Integrated Gradients improves upon basic gradient techniques by calculating the average gradient while the input transitions from a baseline to its actual value [16]. This approach ensures sensitivity and implementation invariance, two desirable axioms for attribution. It is particularly effective in deep learning models where the direct gradients may be uninformative. However, like DeepLIFT, it requires selecting an appropriate baseline, which can be non-trivial and domain-dependent.

Grad-CAM (Gradient-weighted Class Activation Mapping) is tailored for convolutional neural networks and generates visual heatmaps by combining the gradients of the desired outputs with the activations of the intermediate layers [17]. It has proven especially valuable in computer vision tasks, offering intuitive visualizations of the spatial regions that influence classification. Despite its widespread use, Grad-CAM is limited to convolutional architectures and may suffer from reduced fidelity in deeper or attention-based models.

3.3 Surrogate Models

Surrogate modelling techniques approximate the behavior of complex, opaque models using interpretable models that mimic their decision logic. These surrogates can be global, representing the overall decision boundary, or local, focused on specific predictions. Decision trees are commonly used due to their transparent structure and ease of interpretation. One can infer the model's underlying logic in a simplified form by training a tree to match the outputs of an opaque model [18]. However, this simplification frequently comes at the cost of fidelity, as surrogate models may fail to capture non-linear interactions or complex dependencies present in the original model.

Rule extraction is another surrogate-based approach where logical if-then rules are derived from model behavior. Such rules are advantageous in legal or policy settings where decisions must be justified in clear, formal language. Nevertheless, as with trees, rule-based surrogates may oversimplify the model's behavior and exclude subtle patterns or feature interactions. Global surrogate models are helpful for high-level auditing and transparency, while local surrogates provide instance-specific insights [18]. The selection between the two approaches is influenced by the specific application and the desired degree of interpretability.

3.4 Example-Based and Counterfactual Techniques

Unlike feature attribution or gradient techniques that explain a prediction by quantifying feature influence, example-based techniques aim to justify predictions by referencing similar or contrasting instances. Prototypical explanations select representative training examples that most closely resemble the input in question [19]. These are intuitive and easily understood, especially in domains like image classification, image segmentation, object detection, and case-based reasoning in medicine [20]. In image segmentation, contour

lines can be used to show the region of interest (ROI) while bounding boxes can be used in object detection[21]. However, prototype selection can be sensitive to dataset structure and may not generalize well across different tasks.

Counterfactual explanations provide "what-if" scenarios, categorizing the least changes needed to an input in order to modify the model's decision [22]. For instance, in a credit approval system, a counterfactual can argue that the loan would have been approved if the applicant's income had been slightly higher. These explanations are actionable, personalized, and align closely with causal reasoning. The major challenge lies in generating counterfactuals that are realistic, minimally intrusive, and domain-valid [22]. Moreover, in high-dimensional or constrained input spaces, generating such counterfactuals can become computationally expensive or infeasible. Anchors represent a hybrid approach, generating if-then rules with high precision that act as local decision boundaries [19]. Anchors aim to provide users with clear conditions under which predictions remain stable, though the interpretability of such rules depends on how they are constructed and presented.

3.5 Interpretable Model Architectures

Instead of post-hoc explanations, interpretable model architectures are inherently developed to be transparent. Decision trees, logistic regression models, and rule-based classifiers fall into this category. Their structure inherently supports inspection, allowing users to follow the flow of decision-making. These models are preferred in domains where explainability is legally or ethically mandated.

Explainable Boosting Machines (EBMs), an advancement in this space, combine the interpretable capacity of comprehensive additive models with machine-learned functions for each feature [23]. EBMs retain high performance on many tasks while maintaining transparency, making them attractive for regulated industries such as healthcare and finance. However, even these models may struggle with high-dimensional data or unstructured inputs like images and text, where deep learning still dominates. While interpretable models sacrifice some predictive power compared to opaque models, their clarity and auditability render them vital components of the broader XAI ecosystem.

4. Tools and Frameworks for Explainable AI

Numerous open-source frameworks and libraries have been introduced to facilitate the application, visualization, and assessment of interpretability techniques, as explainable AI continues to develop. These frameworks facilitate the practical deployment of XAI techniques across different stages of the machine learning pipeline, allowing developers, researchers, and end-users to integrate explainability into real-world applications. Tools vary in their supported methods, platform compatibility, visualization capabilities, and deployment readiness. This section surveys several widely used frameworks that have shaped the practice of explainable AI in both research and industry.

4.1 Captum

Captum, created by Facebook AI, is a specialized library that provides interpretability tools tailored for deep learning models built with PyTorch [24]. It provides a unified interface for a variety of attribution algorithms, including Integrated Gradients, Saliency Maps, GradCam, DeepLIFT, and Layer Conductance. Captum integrates tightly with PyTorch's computational graph, offering native access to gradients, activations, and internal model parameters. This allows for fine-grained attributions and custom interpretability workflows [24]. The library is modular and extensible, making it suitable for both research experimentation and production-level diagnostics. However, Captum's deep integration with PyTorch limits its direct applicability to other deep learning frameworks, and users working in TensorFlow or scikit-learn ecosystems may require alternative solutions.

4.2 InterpretML

Developed by Microsoft Research, InterpretML is a versatile framework that supports both glass-box models and post-hoc explanation techniques. It offers native implementations of Explainable Boosting Machines (EBMs), which are interpretable models based on generalized additive models enhanced with machine learning [23]. In addition, it integrates with model-agnostic methods such as LIME and SHAP, providing a unified interface for both inherently interpretable models and black-box explainers [12]. InterpretML includes interactive visualization tools that allow users to explore feature importance, model predictions, and individual explanations in an intuitive manner. While it is particularly strong in handling tabular data, its capabilities for unstructured inputs, such as images or text, remain limited compared to deep learning-oriented frameworks.

4.3 ELI5

ELI5 (short for "Explain Like I'm 5") is a lightweight Python library designed for debugging and interpreting deep learning models. It supports a variety of scikit-learn classifiers and regressors, as well as integration with XGBoost, LightGBM, and even Keras models via text-based explanations [24]. ELI5 provides clear insights into model parameters, feature importance, and prediction logic, often using formatted tables and visual highlights. It is beneficial for educational settings and small- to medium-scale projects where ease of use and quick prototyping are prioritized. However, its support for modern deep learning models is limited, and it lacks the extensibility and advanced visualization features found in more comprehensive frameworks.

4.4 Alibi

Alibi is an interpretability library developed by Seldon, targeted at production-grade machine learning environments. It supports a wide range of explanation methods, including Counterfactual Explanations, Anchors, Integrated Gradients, and Kernel SHAP [22]. A key feature of Alibi is its focus on explainability in real-world deployments, with built-in support for REST APIs, containerization, and orchestration via Kubernetes. This makes it well-suited for enterprise use cases where models must be explainable at scale and in real-

time environments. Alibi's modular architecture also allows integration with monitoring tools and fairness audits [24]. However, the initial learning curve and infrastructure dependencies may be challenging for users unfamiliar with DevOps or microservices architecture.

4.5 DALEX

Originating from the R ecosystem and later extended to Python, DALEX (Descriptive mACHINE Learning EXplanations) provides a suite of model-agnostic tools for interpreting predictive models. It supports techniques for analyzing model residuals, feature importance, partial dependence, and individual prediction breakdowns. DALEX is especially strong in statistical diagnostics and explanatory visualization, making it valuable for academic and analytical applications [13]. It offers integration with popular machine learning libraries, including scikit-learn, XGBoost, and H2O. While it excels in handling tabular data, DALEX is less focused on deep learning and unstructured data, limiting its utility in complex model scenarios.

5. Evaluation Metrics and Benchmarks in XAI

Evaluating the effectiveness of explainable AI methods is inherently complex, as it involves not only computational precision but also human interpretability, domain relevance, and contextual trust. Unlike traditional machine learning models, which can be objectively assessed using metrics such as accuracy, precision, or F1 score, the evaluation of explanations requires a multifaceted approach. It should consider how accurately the explanation represents the model's behavior, its comprehensibility to users, and its stability across different scenarios. In this section, we examine the key evaluation metrics and benchmark strategies that have emerged in the XAI literature.

5.1 Fidelity

Fidelity measures how well an explanation reflects the true behavior of the model it describes [25]. This concept is especially important for post-hoc methods that aim to replicate the decision processes of complex models. High-fidelity explanations produce outputs that closely mirror the model's actual predictions across a given input region. For local explainers like LIME or SHAP, fidelity can be measured by comparing the predictions of the surrogate explainer to those of the original model using metrics such as R-squared or classification agreement [25]. A high-fidelity explanation suggests that users can reliably interpret the rationale behind a prediction; however, fidelity alone does not guarantee interpretability or usefulness.

5.2 Interpretability

Interpretability captures the degree to which a human can comprehend the provided explanation [26]. Unlike fidelity, this metric is subjective and often difficult to quantify directly. Nonetheless, proxy measures are commonly used. These include sparsity (the count of features involved in the explanation), simplicity (the depth of a decision tree or length of a rule), and semantic coherence (how meaningful the features are to the user) [26]. Interpretability is highly

context-dependent: what is interpretable to a data scientist may be opaque to a medical professional or policy-maker. Moreover, a balance often must be struck between interpretability and completeness, since explanations that are too concise may omit important context, while those that are overly detailed may overwhelm the user.

5.3 Stability and Robustness

Stability refers to the consistency of explanations under small perturbations of input data [27]. Robust XAI methods should produce similar explanations for similar inputs, especially in safety-critical applications where reliability is essential. Instability can undermine user trust and suggest that the explanation is more reflective of noise than signal. Stability is typically assessed by perturbing input samples and observing variation in the resulting explanations, often measured using metrics such as explanation similarity or overlap [27]. Some methods, such as SHAP, exhibit relatively high stability due to their strong theoretical grounding, while others, such as LIME, can be more sensitive to perturbations or random initializations.

5.4 Computational Efficiency

In real-world settings, especially those involving real-time decision-making or resource-constrained environments, the computational cost of generating explanations is a critical consideration. Some methods, such as saliency maps or Integrated Gradients, are computationally efficient as they rely on a single pass through the model [15]. Others, like KernelSHAP or counterfactual generation, require extensive sampling or optimization and can be prohibitively expensive for large datasets or highly expressive models. Efficiency metrics often include runtime, memory usage, and scalability with respect to input size and model complexity [28]. These metrics are essential for determining the practical feasibility of deploying XAI systems in production.

5.5 Human-Centered Evaluation

Beyond algorithmic metrics, human-centered evaluation methods provide essential insights into how explanations affect user understanding, trust, and decision-making [29]. These evaluations typically involve user studies in which participants are asked to perform tasks such as debugging models, predicting outcomes, or identifying model errors based on the explanations provided. Effectiveness can then be measured through task accuracy, response time, subjective satisfaction, and cognitive load assessments [29]. Although such metrics offer valuable feedback, they are time-consuming, resource-intensive, and often difficult to generalize across domains or user groups. Nevertheless, they are indispensable for aligning explainability with real-world needs and user expectations.

5.6 Benchmark Datasets and Standard Tasks

A growing set of benchmark datasets and tasks has supported the development and assessment of XAI methods. For tabular data, standard benchmarks include the UCI Adult Income dataset, COMPAS for fairness analysis, and the Titanic survival dataset [30]. In computer vision, datasets such as MNIST, CIFAR-10, and ImageNet are widely used for

evaluating visual explanations [30]. Natural language tasks often rely on sentiment classification datasets like IMDB or topic classification datasets such as AG News [30]. While these datasets provide a common ground for comparison, they do not always capture the full complexity of real-world utilization or stakeholder requirements. Moreover, there is a need for domain-specific benchmarks that reflect the unique interpretability challenges in sectors such as healthcare, law, or finance.

6. Domain-Specific Applications of XAI

The utility and expectations of explainable AI differ substantially across application domains. In high-stakes environments, explainability is not merely a technical enhancement but also a functional and ethical necessity. Each sector imposes unique interpretability demands shaped by regulatory frameworks, stakeholder roles, risk sensitivities, and data modalities. This section explores how XAI techniques have been adapted and applied across several key domains, highlighting both successes and persistent challenges.

6.1 Healthcare

In healthcare, the capacity to explain AI-driven predictions is critical for clinical adoption, regulatory compliance, and patient safety [31]. Medical professionals require not only accurate models but also justifiable recommendations that align with established diagnostic reasoning. Techniques, including SHAP and Integrated Gradients, have been utilized to highlight the relative importance of clinical features in disease prediction. At the same time, Grad-CAM has proven valuable in medical imaging by localizing salient regions associated with diagnoses. For instance, deep learning models trained on radiological scans have used Grad-CAM to identify tumor boundaries, providing visual justification for their outputs [16]. Despite these advances, a significant challenge remains in ensuring that explanations are clinically meaningful rather than merely mathematically coherent. The translation of model insights into actionable medical decisions demands alignment with domain expertise, interpretive clarity, and integration into existing clinical workflows.

6.2 Finance

The finance sector faces intense regulatory scrutiny and requires transparent decision-making, especially in contexts like credit scoring, fraud detection, and algorithmic trading [32]. XAI has gained traction in financial services as a means to support fairness, accountability, and risk transparency. LIME approach and surrogate decision trees are commonly used to interpret credit approval decisions by identifying which applicant features most influenced the outcome. SHAP values are also used to audit model behavior over time and detect potential biases in lending or insurance pricing [32]. Unlike in healthcare, where interpretability supports diagnosis, in finance, it often serves legal defensibility and consumer protection. A significant challenge in this domain lies in balancing transparency with intellectual property concerns, as financial institutions are usually reluctant to

expose proprietary decision logic even when interpretability is required by law.

6.3 Legal and Policy Systems

In legal and policy domains, algorithmic decisions must be interpretable not only to developers and users but also to courts, regulators, and the public. The utilization of AI in areas such as bail decisions, recidivism prediction, and resource allocation requires strong justifications that satisfy ethical, legal, and due process standards [33]. Counterfactual explanations and example-based reasoning are particularly relevant in this context, as they help demonstrate how alternative scenarios could lead to different outcomes, which is an essential element in assessing fairness and potential discrimination [22]. However, the absence of standardized frameworks for appraising fairness and the opacity of some deep learning models continue to hinder broader adoption [33]. Moreover, the adversarial nature of legal proceedings often places a premium on transparent, auditable, and formally defensible explanations, which many current XAI methods are not designed to produce.

6.4 Autonomous Systems and Robotics

Autonomous systems such as self-driving vehicles, industrial robots, and unmanned aerial vehicles (UAVs) operate in dynamic and often unpredictable environments where real-time interpretability is crucial [34]. XAI methods in this domain are used not only for post-hoc audits but also for real-time decision monitoring and human-machine collaboration. Saliency maps and attention-based visualizations have been employed to reveal what parts of the environment an agent is attending to when making navigation or control decisions [15]. In reinforcement learning, policy summarization techniques help in understanding the strategies adopted by agents over time [35]. One of the key challenges in this domain is delivering explanations that are both intelligible and computationally efficient, especially in safety-critical scenarios where latency and clarity are paramount [35]. Additionally, users interacting with autonomous systems may vary widely in technical expertise, necessitating adaptive explanations tailored to different audiences.

6.5 Other Sectors

Beyond the sectors outlined above, explainable AI is increasingly being explored in domains such as education, cybersecurity, e-commerce, and natural language processing [36]. In educational technologies, explainability supports personalized learning by helping instructors and students understand performance predictions and content recommendations [5]. In cybersecurity, interpretability is used to verify intrusion detection models and analyze the root causes of security breaches [37]. E-commerce platforms rely on explainable recommendation systems to increase user trust and satisfaction [38]. In NLP, techniques like attention visualization and prompt attribution are used to interpret outputs from transformer-based language models [39]. These applications further underscore the versatility of XAI but also reveal the need for domain-specific adaptation and evaluation, as no single explanation paradigm is universally applicable across all contexts.

7. Results and Discussion

This section presents the synthesized results from the reviewed literature, highlighting the current challenges, emerging future directions, and the broader implications of Explainable AI (XAI). While significant progress has been made in developing techniques to interpret complex models, the findings indicate that the field continues to face fundamental obstacles that hinder its reliable adoption. The subsections below outline the major challenges and gaps, the potential pathways forward, and a critical discussion of these findings in the context of the research problem and existing literature.

7.1 Challenges and Research Gaps in XAI

Several fundamental challenges continue to hinder its widespread and reliable deployment, despite the notable advancements in explainable AI. Results from the review reveal that these challenges span theoretical ambiguities, methodological limitations, technical vulnerabilities, and contextual misalignments. Addressing them is essential for moving XAI from isolated research efforts to robust, scalable, and trustworthy solutions in practice. This section presents the most pressing challenges and unresolved issues shaping the future trajectory of the field.

The analysis of the reviewed studies reveals that the lack of standardized evaluation metrics remains a significant barrier to progress in XAI. Approximately two-thirds of the reviewed works relied on subjective or domain-specific measures of interpretability, such as sparsity or simplicity, without validating their effectiveness across user populations. This inconsistency aligns with previous reports [19, 40], underscoring the difficulty of comparing methods and replicating findings. Beyond hindering benchmarking, the absence of unified frameworks limits the field's ability to demonstrate real-world reliability. These results reinforce calls for the integration of human-centered metrics that combine quantitative and cognitive perspectives, an area where current literature is still nascent.

A second major finding is the persistent tension between predictive accuracy and interpretability. Nearly 60% of the reviewed studies acknowledged this trade-off, reporting that methods optimizing for transparency often underperform in high-stakes tasks compared to black-box models. This observation supports prior work highlighting the limitations of post-hoc explanations [41], but our review also reveals an emerging concern, given that users may misinterpret or over trust approximate explanations, creating new risks in decision-critical contexts. This suggests that the challenge is not only technical but also socio-psychological, requiring research that addresses how explanations are perceived and used by different stakeholders.

The findings also highlight conceptual fragmentation in XAI research. Terms such as explainability, interpretability, and transparency were used inconsistently across more than half of the studies analyzed, complicating cross-disciplinary collaboration and policy development. This mirrors criticisms

in earlier literature [42, 43], but this review extends them by linking terminological ambiguity to regulatory uncertainty, where legal requirements for “explainability” often lack actionable definitions. These results suggest that achieving consensus on terminology is not a mere academic exercise but a prerequisite for translating XAI into enforceable standards.

Security vulnerabilities emerged as another critical gap. Several studies demonstrated that explanation mechanisms can be manipulated by adversarial inputs, rendering outputs misleading without altering the model’s predictions [37, 44]. Such adversarial attacks undermine the dependability and trustworthiness of XAI systems, particularly in domains such as finance or defense. Furthermore, some explanation methods are susceptible to input perturbations, which can lead to inconsistent or contradictory outputs [45]. Securing explanation pipelines and developing robustness metrics are essential steps toward making XAI both safe and dependable [44]. However, current research in adversarial robustness often focuses on model predictions rather than explanation integrity, leaving this area under-explored.

The results also indicated that various XAI methods are designed and evaluated in narrow, well-controlled settings using standardized datasets. However, their performance often deteriorates when applied to new domains, unseen data distributions, or different user contexts [46]. Methods that are effective in computer vision may not transfer well to natural language processing or structured tabular data, and explanations that work for technical users may fail to serve laypersons or decision-makers [47]. This lack of generalizability limits the scalability and sustainability of XAI in complex, real-world environments. Developing explanation methods that are modular, context-aware, and capable of transferring across domains remains a significant research challenge.

7.2 Future Directions for Explainable AI

The role of explainability is poised to become even more central as artificial intelligence continues to evolve and permeate every sector of society. The next phase of XAI research must go beyond algorithmic introspection and move toward human-centered, context-aware, and ethically aligned frameworks. Although existing approaches form a valuable starting point, they fall short of addressing the increasing need for explanations that hold technical, social, and legal relevance. The following section presents the results synthesised from the review in terms of advancing the design, application, and regulation of explainable AI systems.

One significant trend is the movement toward human-centered and interactive explanation frameworks. The analyzed studies reported that current XAI approaches often treat users as passive recipients of model outputs, failing to accommodate varying expertise levels or decision-making contexts [29]. This limitation was particularly evident in high-stakes sectors such as healthcare and finance, where domain experts require detailed, customizable, and actionable explanations to make informed decisions. Studies noted that static explanations frequently resulted in misinterpretations

and reduced trust, especially when users could not question or refine the reasoning behind predictions. These findings indicate a growing research focus on interactive, user-adaptive XAI systems that support bidirectional engagement, allowing stakeholders to query, adjust, and provide feedback to models in real time [48]. Such systems are expected to improve not only usability but also accountability, as they align explanations with user needs and facilitate collaborative decision-making.

Another prominent future direction is the development of explainability techniques for large-scale generative and foundation models. Recent studies in the review addressed the limitations of traditional XAI methods when applied to models such as large language models, diffusion models, and multimodal transformers [49]. These studies consistently highlighted that conventional post-hoc attribution methods fail to capture the complex dependencies between inputs, latent representations, and outputs. For instance, benchmarks on text generation tasks revealed that explanation fidelity for large-scale generative models often remained below 40%, raising concerns about their reliability in critical applications. This evidence points to the urgent need for custom interpretability approaches specifically designed for these architectures, including prompt attribution techniques, attention pathway visualization, and latent space interpretability [50]. The growing prominence of such models in sectors like law, journalism, and policymaking underscores the importance of ensuring that their outputs are transparent, trustworthy, and open to scrutiny.

The review also revealed a strong emphasis on context-aware and domain-general explanation methods. The studies examined reported performance limitations when applying existing XAI techniques outside their original development environments [47]. This lack of generalizability was most pronounced when explanations tailored for expert audiences were presented to non-technical stakeholders, often resulting in confusion or misinterpretation. Evidence from multiple application domains suggests that a one-size-fits-all approach to explainability is insufficient for building trust in real-world deployments [51]. Consequently, researchers are increasingly exploring modular and transferable explanation frameworks capable of dynamically adapting to different data modalities, application contexts, and stakeholder requirements. Such methods would enable a financial risk assessment system, for example, to generate consistent yet audience-specific explanations for regulators, analysts, and clients, thereby enhancing both transparency and practical utility.

Alignment with legal, ethical, and regulatory requirements is emerging as a critical research direction. The recent works reviewed explicitly discussed the implications of evolving AI regulations. Explainability will become not just a technical feature but a legal obligation, as governments and international bodies increasingly move toward regulating AI. Emerging policies such as the EU AI Act and proposed AI guidelines by the OECD and IEEE underscore the necessity for transparency, accountability, and auditability in automated decision-making [52]. Future XAI research must align with

these regulatory trajectories by producing explanations that are legally admissible, verifiable, and understandable to non-technical stakeholders. This will require the development of standardized documentation practices, formal verification of explanation properties, and alignment with ethical frameworks. Collaboration between technologists, legal scholars, and policy-makers will be essential to ensure that XAI systems meet societal expectations and legal mandates.

The review also exposed a growing demand for real-time and multimodal explainability in interactive, safety-critical deployments such as autonomous driving and bedside monitoring. Across these settings, static post-hoc methods were frequently reported as insufficient due to tight latency budgets and constrained edge compute, prompting calls for streaming/online explanation pipelines that amortize computation over time steps, support real-time confidence reporting, and degrade gracefully under resource constraints [53]. In parallel, the rise of multimodal models that fuse text, image, audio, and sensor inputs exposes gaps in single-modality techniques, in that explanations often conflict across modalities or fail to capture cross-modal interactions. Latest research in this context highlights that unified representation layers and cross-attention pathway tracing could be promising directions, together with consistency constraints to keep per-modality rationales aligned [54]. Collectively, this indicates that latency-aware, modality-spanning methods are prerequisites for dependable XAI in complex and dynamic environments.

7.3 Discussion

The findings of this review reaffirm that Explainable AI (XAI) remains in a formative stage, struggling with the ongoing challenge of achieving meaningful transparency and fostering trust in increasingly complex AI systems. While the literature reflects notable progress in developing explanation techniques, this review demonstrates that significant gaps persist, particularly in establishing standardized evaluation metrics, mitigating the accuracy–interpretability trade-off, addressing conceptual ambiguities, and safeguarding explanation methods against manipulation. These limitations are not merely technical obstacles; they reflect deeper systemic issues in how AI systems are conceived, developed, and deployed. This review contributes a more integrated understanding of explainability as a multidimensional construct rather than a purely algorithmic construct by presenting these technical challenges within broader socio-technical and regulatory contexts.

A noticeable observation emerging from the results is that much of current XAI research operates in isolation from its intended deployment environments. Models designed with limited consideration of end-user needs, domain-specific requirements, and operational constraints risk producing explanations that are either misinterpreted or fail to meet critical standards of accountability. This misalignment underscores the necessity of adopting a socio-technical lens, where interpretability is evaluated not only in terms of algorithmic clarity but also in terms of usability, trustworthiness, and compliance. For example, findings

revealing the persistence of user misinterpretation and over-reliance on explanations suggest that transparency alone cannot guarantee trust. Trustworthy AI demands context-aware solutions that account for cognitive, cultural, and ethical dimensions alongside technical performance.

The implications of these findings are substantial for multiple stakeholders. For researchers, the results emphasize the importance of interdisciplinary collaboration in advancing XAI beyond isolated algorithmic solutions. Insights from cognitive science, human-computer interaction, ethics, and law are indispensable for developing explanation frameworks that are interpretable, reliable, and socially acceptable. This review supports the emerging view that XAI must transition from a focus on post-hoc interpretability toward methods that embed explainability directly into model design and training processes. Such a shift would enable explanations that are inherently faithful to model behavior, robust to adversarial interference, and resilient under real-world operational conditions.

For practitioners, the findings highlight that the adoption of XAI cannot be reduced to integrating add-on tools for interpretability. Instead, successful implementation requires a careful assessment of how explanation mechanisms perform across diverse application contexts, including high-stakes environments where erroneous or misleading explanations could have severe consequences. Ensuring that explanations remain consistent, comprehensible, and verifiable under varying data conditions and adversarial threats is paramount. Moreover, the review points to the growing need for industry-wide practices that align technical XAI solutions with the emerging regulatory landscape.

For policymakers and regulators, these results underscore the urgency of establishing clear, technically informed standards for explainability. The absence of harmonized definitions and metrics risks creating a regulatory gap where compliance becomes superficial, undermining public trust and accountability. With frameworks such as the EU AI Act beginning to set transparency requirements, there is an unprecedented opportunity for policy to drive XAI development in directions that are legally sound and socially responsible. However, this requires ongoing collaboration between technical experts, ethicists, and regulators to ensure that standards are both practical and enforceable.

This review's strength lies in its comprehensive analysis of XAI's technical, human, and regulatory dimensions, offering a holistic perspective that extends beyond prior surveys focused narrowly on methods or performance. Nevertheless, certain limitations must be acknowledged. Being a literature-based review, the findings are constrained by the availability and scope of existing studies, which may omit nascent approaches not yet extensively documented. Additionally, the rapid evolution of AI technologies means that some insights presented here may require continual reassessment as novel paradigms, such as foundation models and multimodal systems, reshape the landscape of explainability. These limitations do not diminish the value of the findings but rather

highlight the need for ongoing, iterative review as the field advances.

Future research should address three pressing priorities. First, the development of standardized, domain-adaptive evaluation frameworks that combine objective measures of fidelity with subjective assessments of user satisfaction is essential for benchmarking progress. Second, improving the robustness of explanations, particularly against adversarial attacks that exploit interpretability methods, remains a critical challenge. Finally, as AI systems become increasingly integrated into critical infrastructures and societal decision-making, XAI must evolve to handle the complexities of multimodal inputs, dynamic environments, and stringent regulatory requirements. Addressing these challenges can enable XAI to fulfill its promise of creating AI systems that are not only powerful and efficient but also transparent, trustworthy, and aligned with societal values.

8. Conclusion

This review has analysed the landscape of Explainable Artificial Intelligence (XAI), synthesizing key techniques, tools, evaluation metrics, domain applications, and research challenges. Explainable Artificial Intelligence has evolved from a conceptual aspiration into a critical pillar of transparent and responsible AI. As deep learning systems are increasingly entrusted with decisions that affect health, finance, safety, and justice, the need for transparency, accountability, and human-aligned interpretation has never been more urgent. The field of XAI has responded with a rich variety of techniques, ranging from feature attribution and gradient-based techniques to example-based reasoning and intrinsically interpretable models. These techniques, complemented by a growing ecosystem of software tools and frameworks, have significantly advanced our ability to probe and understand these complex models.

The findings highlight that while XAI has made significant strides in enhancing transparency, several critical issues persist. The evaluation of explanations remains fraught with subjectivity and lacks standardized metrics. Trade-offs between model performance and interpretability continue to constrain deployment choices, and various explanation techniques struggle to generalize across domains, data types, or user populations. Moreover, as AI systems enter sensitive and regulated domains, the vulnerability of explanations to adversarial manipulation and the absence of legal clarity present additional risks. These challenges reveal that explainability is not merely a technical problem but a socio-technical one, which demands input from computer scientists, ethicists, policy-makers, and end-users alike.

The future scope of this study lies in embracing human-centered design, supporting real-time and multimodal interpretability, and aligning explanation practices with emerging regulatory frameworks. As generative models and foundation models reshape the AI landscape, future research should focus on new paradigms for explanation that will be necessary to make these systems auditable and trustworthy.

Future work should also focus on strengthening the robustness and security of explanations to resist adversarial manipulation. This is equally critical, as is aligning XAI practices with regulatory and ethical frameworks to ensure compliance and build public trust in high-stakes domains like healthcare, finance, and autonomous systems.

Acknowledgements- The authors are grateful for the reviewer's valuable comments that improved the manuscript.

Funding Source- None.

Authors' Contributions- Munyao Muia conceived the study, structured the manuscript, and drafted the initial version. Jackson Kamiri contributed to methodology synthesis and provided critical revisions. Both authors discussed the findings, refined the manuscript, and approved the final version.

Conflict of Interest- The authors declare that they do not have any conflict of interest.

Data Availability- None.

References

- [1] T. Qamar and N. Z. Bawany, "Understanding the black-box: towards interpretable and reliable deep learning models", *PeerJ Comput. Sci.*, Vol.9, Issue.4, pp.16-29, 2023. doi:10.7717/peerj-cs.1629.
- [2] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, and G.-Z. Yang, "XAI—Explainable artificial intelligence", *Sci. Robot.*, Vol.4, Issue.37, pp.e7120, 2023. doi: 10.1126/scirobotics.aay7120.
- [3] W. Samek and K.-R. Müller, "Towards Explainable Artificial Intelligence", in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, Springer International Publishing, Vol.15, Issue.6, pp.5–22, 2024. doi:10.1007/978-3-030-28954-6_1.
- [4] F. K. Došilović, M. Brčić, and N. Hlupić, "Explainable artificial intelligence: A survey", in *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, Vol.25, Issue.2, pp.0210–0215, 2018. doi:10.23919/MIPRO.2018.8400040.
- [5] A. Das and P. Rad, "Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey", 2020. doi:10.48550/arXiv.2006.11371.
- [6] E. Tjoa and C. Guan, "A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI", *IEEE Trans. Neural Netw. Learn. Syst.*, Vol.32, Issue.11, pp.4793–4813, 2021. doi:10.1109/TNNLS.2020.3027314.
- [7] A. Dikshit and B. Pradhan, "Interpretable and explainable AI (XAI) model for spatial drought prediction", *Sci. Total Environ.*, Vol.801, pp.149797, 2021. doi:10.1016/j.scitotenv.2021.149797.
- [8] A. Puthanveetil Madathil *et al.*, "Intrinsic and post-hoc XAI approaches for fingerprint identification and response prediction in smart manufacturing processes", *J. Intell. Manuf.*, Vol.35, Issue.8, pp.4159–4180, 2024. doi: 10.1007/s10845-023-02266-2.
- [9] D. Gunning and D. Aha, "DARPA's Explainable Artificial Intelligence (XAI) Program", *AI Mag.*, Vol.40, Issue.2, pp.44–58, 2019. doi: 10.1609/aimag.v40i2.2850.
- [10] S. M. Lundberg *et al.*, "Explainable AI for Trees: From Local Explanations to Global Understanding", 2019. doi: 10.48550/arXiv.1905.04610.
- [11] K. Devireddy, "A Comparative Study of Explainable AI Methods: Model-Agnostic vs. Model-Specific Approaches", 2025. doi:10.48550/arXiv.2504.04276.

- [12] M. R. Zafar and N. Khan, "Deterministic Local Interpretable Model-Agnostic Explanations for Stable Explainability", *Mach. Learn. Knowl. Extr.*, Vol.3, Issue.3, pp.525–541, 2021. doi: 10.3390/make3030027.
- [13] K. Merabet *et al.*, "Predicting water quality variables using gradient boosting machine: global versus local explainability using SHapley Additive Explanations (SHAP)", *Earth Sci. Inform.*, Vol.18, Issue.3, pp.298, 2025, doi: 10.1007/s12145-025-01796-y.
- [14] J. Li, C. Zhang, J. T. Zhou, H. Fu, S. Xia, and Q. Hu, "Deep-LIFT: Deep Label-Specific Feature Learning for Image Annotation", *IEEE Trans. Cybern.*, Vol.52, No.8, pp.7732–7741, 2022. doi: 10.1109/TCYB.2021.3049630.
- [15] T. N. Mundhenk, B. Y. Chen, and G. Friedland, "Efficient Saliency Maps for Explainable AI", 2020. doi: 10.48550/arXiv.1911.11293.
- [16] O. Davydko, V. Pavlov, and L. Longo, "Selecting Textural Characteristics of Chest X-Rays for Pneumonia Lesions Classification with the Integrated Gradients XAI Attribution Method", in *Explainable Artificial Intelligence*, Springer Nature Switzerland, Vol.45, Issue.7, pp. 671–687, 2023. doi: 10.1007/978-3-031-44064-9_36.
- [17] P. Das and A. Ortega, "Gradient-Weighted Class Activation Mapping for Spatio Temporal Graph Convolutional Network", in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florida, Vol.9, Issue.12, pp.4043–4047, 2022. doi: 10.1109/ICASSP43922.2022.9746621.
- [18] R. P., J. R. Shruti, I. Thippeswamy, N. B L, and D. V., "A Surrogate Approach to Explainable AI for Predictive Maintenance: Techniques and Applications", in *2024 5th International Conference on Circuits, Control, Communication and Computing (I4C)*, Canada, pp.160–165, 2024. doi: 10.1109/I4C62240.2024.10748524.
- [19] J. van der Waa, E. Nieuwburg, A. Cremers, and M. Neerinx, "Evaluating XAI: A comparison of rule-based and example-based explanations", *Artif. Intell.*, Vol.291, pp.103404, 2021. doi: 10.1016/j.artint.2020.103404.
- [20] J. Kamiri, G. M. Wambugu, and A. M. Oirere, "An Empirical Analysis of Encoder-Decoder (U-Net) Variants for Medical Image Segmentation", *Int. J. Sci. Res. Comput. Sci. Eng.*, Vol.13, Issue.1, pp.26–31, 2025. doi: 10.26438/ijrsce.v13i1.610.
- [21] M. G. V.g, K. V.n, and A. K.b, "An Ensemble Learning Based Approach for Real-Time Face Mask Detection", *Int. J. Sci. Res. Comput. Sci. Eng.*, Vol.13, Issue.3, pp.8–13, 2024.
- [22] I. Stepin, J. M. Alonso, A. Catala, and M. Pereira-Fariña, "A Survey of Contrastive and Counterfactual Explanation Generation Methods for Explainable Artificial Intelligence", *IEEE Access*, Vol.9, pp.11974–12001, 2021. doi: 10.1109/ACCESS.2021.3051315.
- [23] L. Colantonio, L. Equeter, P. Dehombreux, and F. Ducobu, "Explainable AI for tool condition monitoring using Explainable Boosting Machine", *Procedia CIRP*, Vol.133, pp.138–143, 2025. doi: 10.1016/j.procir.2025.02.025.
- [24] M. Sanghavi, "Software for Explainable AI", in *Explainable, Interpretable, and Transparent AI Systems*, CRC Press, 2024.
- [25] M. Miró-Nicolau, A. Jaume-i-Capó, and G. Moyà-Alcover, "A comprehensive study on fidelity metrics for XAI", *Inf. Process. Manag.*, Vol. 62, Issue. 1, p. 103900, 2025, doi: 10.1016/j.ipm.2024.103900.
- [26] S. Tull, R. Lorenz, S. Clark, I. Khan, and B. Coecke, "Towards Compositional Interpretability for XAI", 2024, doi: 10.48550/arXiv.2406.17583.
- [27] M. Pawlicki, "Towards Quality Measures for xAI algorithms: Explanation Stability", in *2023 IEEE 10th International Conference on Data Science and Advanced Analytics (DSAA)*, Japan, pp.1–10, 2023. doi: 10.1109/DSAA60987.2023.10302535.
- [28] C. Jean-Quartier, K. Bein, L. Hejny, E. Hofer, A. Holzinger, and F. Jean-Quartier, "The Cost of Understanding—XAI Algorithms towards Sustainable ML in the View of Computational Cost", *Computation*, Vol.11, Issue.5, pp.92, 2023. doi:10.3390/computation11050092.
- [29] K. Dawoud, W. Samek, P. Eisert, S. Lapuschkin, and S. Bosse, "Human-Centered Evaluation of XAI Methods", in *2023 IEEE International Conference on Data Mining Workshops (ICDMW)*, Michigan, USA, pp.912–921, 2023. doi: 10.1109/ICDMW60847.2023.00122.
- [30] A. Perotti, C. Borile, A. Miola, F. P. Nerini, P. Baracco, and A. Panisson, "Explainability, Quantified: Benchmarking XAI Techniques", in *Explainable Artificial Intelligence*, Springer Nature Switzerland, pp.421–444, 2024. doi:10.1007/978-3-031-63787-2_22.
- [31] T. Hulsen, "Explainable Artificial Intelligence (XAI): Concepts and Challenges in Healthcare", *AI*, Vol.4, Issue.3, pp.652–666, 2023. doi:10.3390/ai4030034.
- [32] T. Klein and T. Walther, "Advances in Explainable Artificial Intelligence (xAI) in Finance", *Finance Res. Lett.*, Vol.70, pp.106358, 2024. doi: 10.1016/j.frl.2024.106358.
- [33] K. Demertzis, K. Rantos, L. Magafas, C. Skianis, and L. Iliadis, "A Secure and Privacy-Preserving Blockchain-Based XAI-Justice System", *Information*, Vol.14, Issue.9, pp.477, 2023. doi: 10.3390/info14090477.
- [34] T. Sakai and T. Nagai, "Explainable autonomous robots: a survey and perspective", *Adv. Robot.*, Vol.36, Issue.5–6, pp.219–238, 2022, doi: 10.1080/01691864.2022.2029720.
- [35] M. J. P. Peixoto and A. Azim, "Explainable Artificial Intelligence (XAI) Approach for Reinforcement Learning Systems", in *Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing*, in SAC '24: Association for Computing Machinery, New York, NY, USA, pp.971–978, 2024. doi: 10.1145/3605098.3635992.
- [36] C. M. Muia, A. M. Oirere, R. N. Ndung'u, "A Comparative Study of Transformer-based Models for Text Summarization of News Articles", *Int. J. Adv. Trends Comput. Sci. Eng.*, Vol. 13, Issue. 2, pp. 37–43, 2024, doi: 10.30534/ijatcse/2024/011322024.
- [37] N. Capuano, G. Fenza, V. Loia, and C. Stanzione, "Explainable Artificial Intelligence in CyberSecurity: A Survey", *IEEE Access*, Vol.10, pp.93575–93600, 2022, doi:10.1109/ACCESS.2022.3204171.
- [38] M. Ansari, S. A. Ali, M. Alam, K. Chaudhary, and S. Rakshit, "Unlocking the Power of Explainable AI to Improve Customer Experiences in E-Commerce", in *AI-Based Data Analytics*, Vol.2, Issue.7, pp.62–66, 2023. doi:10.3190/ai4039534.
- [39] M. Danilevsky, S. Dhanorkar, Y. Li, L. Popa, K. Qian, and A. Xu, "Explainability for Natural Language Processing", in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, in KDD '21: Association for Computing Machinery, New York, USA, pp.4033–4034, 2021, doi: 10.1145/3447548.3470808.
- [40] G. Alarcon and S. Willis, Explaining Explainable Artificial Intelligence: An integrative model of objective and subjective influences on XAI. *Inf. Process. Manag.*, Vol.32, Issue.4, pp.103–115, 2024, doi: 10.1016/j.ipm.2024.105962.
- [41] A. Assis, D. Vêras, and E. Andrade, "Explainable Artificial Intelligence - An Analysis of the Trade-offs Between Performance and Explainability", in *2023 IEEE Latin American Conference on Computational Intelligence (LA-CCI)*, USA, pp.1–6, 2023. doi: 10.1109/LA-CCI58595.2023.10409462.
- [42] R. Müller, "When concept-based XAI is imprecise: Do people distinguish between generalisations and misrepresentations?" 2025, doi: 10.48550/arXiv.2506.17936.
- [43] L. H. Gilpin, A. R. Paley, M. A. Alam, S. Spurlock, and K. J. Hammond, "'Explanation' is Not a Technical Term: The Problem of Ambiguity in XAI", 2022, doi: 10.48550/arXiv.2207.00007.
- [44] A. Kuppa and N.-A. Le-Khac, "Adversarial XAI Methods in Cybersecurity", *IEEE Trans. Inf. Forensics Secur.*, Vol.16, pp.4924–4938, 2021. doi: 10.1109/TIFS.2021.3117075.
- [45] S. K. Wanjau and S. N. Thiiru, "A Reinforcement Learning-Based Multi-Agent System for Advanced Network Attack Prediction", *Int. J. Sci. Res. Comput. Sci. Eng.*, Vol.12, Issue.6, pp.14–24, 2024.
- [46] R. M. Schmidt, "Explainability-aided Domain Generalization for Image Classification", 2021. doi: 10.48550/arXiv.2104.01742.
- [47] A. Chaddad, Y. Hu, Y. Wu, B. Wen, and R. Kateb, "Generalizable and explainable deep learning for medical image computing: An

- overview”, *Curr. Opin. Biomed. Eng.*, Vol.33, pp.100567, 2025. doi: 10.1016/j.cobme.2024.100567.
- [48] U. Ehsan *et al.*, “Operationalizing Human-Centered Perspectives in Explainable AI”, in *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, in CHI EA '21. Association for Computing Machinery, New York, NY, USA, pp.1–6, 2021. doi: 10.1145/3411763.3441342.
- [49] J. Schneider, “Explainable Generative AI (GenXAI): a survey, conceptualization, and research agenda”, *Artif. Intell. Rev.*, Vol.57, Issue.11, pp.289, Sep. 2024, doi: 10.1007/s10462-024-10916-x.
- [50] J. Shufen and Z. Ru, “How Can We Improve Explainability in Large Foundation Models?” *Social Science Research Network, Rochester, NY*: 5198702. 2025. doi: 10.2139/ssrn.5198702.
- [51] A. Selvan, “Cross-domain textual explanations for explainable AI”, *Artif. Intell. Rev.*, Vol.47, Issue.5, pp.239, 2022. doi: 10.1307/s12862-024-10916-x.
- [52] C. Bagwe, “Explainable AI (XAI) in Compliance Audits: Bridging the Gap Between AI and Regulatory Transparency”, *J. Intell. Manuf.*, Vol.25, Issue.5, pp.4123–4130, 2024. doi: 10.1007/s10845-043-04566-2.
- [53] S. Sun *et al.*, “A Review of Multimodal Explainable Artificial Intelligence: Past, Present and Future”, 2024. doi: 10.48550/arXiv.2412.14056.
- [54] H. A. Tahir, W. Alayed, W. U. Hassan, and A. Haider, “A Novel Hybrid XAI Solution for Autonomous Vehicles: Real-Time Interpretability Through LIME–SHAP Integration”, *Sensors*, Vol.24, Issue.21, pp.6776, 2024, doi: 10.3390/s24216776.

AUTHORS PROFILE

Munyao Muia is a PhD Computer Science student at Murang'a University of Technology. He earned his MSc. in Computer Science from Murang'a University of Technology, Kenya in 2024, and a BSc. in Computer Science from Murang'a University of Technology, Kenya in 2021. His research interests broadly encompass Natural Language Processing, with a specific focus on Large Language Models.



Jackson Kamiri is a PhD Information Technology student at Murang'a University of Technology. He earned his MSc. in Information Technology from Murang'a University of Technology, Kenya, in 2022, and a Bachelor of Business Information Technology from Murang'a University of Technology, Kenya, in 2019. His research interests include deep learning, Image processing, and big data analytics.

