ISSN - 2456-1142

International journal of humanities and information technology Volume 8, Issue 3, 2026

Machine Learning Model Interpretability and Explainability

Opeyemi Ajayi

Obafemi Awolowo university ile ife

Abstract

The increasing deployment of complex machine learning models in high-stakes domains has created an urgent need for interpretability and explainability. This study investigates the effectiveness of various explainable artificial intelligence (XAI) techniques across three critical dimensions: faithfulness, stability, and comprehensibility. Using a mixed-methods design combining quantitative evaluation with user-centric assessment, we examined five XAI methods—LIME, SHAP, Integrated Gradients, Layer-wise Relevance Propagation (LRP), and Attention Mechanism Visualization (AMV)—across multiple model architectures and two downstream tasks. Quantitative results demonstrate that LIME consistently achieves the highest Human-reasoning Agreement (MAP = 0.82), while AMV shows superior robustness with a Mean Average Difference of 0.24 against input perturbations. LRP excels in contrastivity, particularly with complex transformer models. However, user studies with 39 participants across three expertise groups revealed that technical accuracy of explanations does not guarantee user comprehension, with non-experts understanding only 58% of explanation content compared to 89% for experts. Qualitative findings indicate that domain knowledge integration significantly improves explanation quality, with semantically structured representations improving alignment with expert reasoning by 43%. These findings contribute to a nuanced understanding of XAI effectiveness and provide practical guidelines for designing interpretable machine learning systems that balance technical fidelity with human comprehensibility.

Keywords: Interpretable Machine Learning, Explainable AI, Faithfulness, Stability, Human-centered AI

International journal of humanities and information technology (2026)

Introduction

Background and Context

Machine learning models have achieved remarkable success across diverse domains, from healthcare and autonomous driving to finance and scientific discovery. However, the most accurate models are often based on complex "black-box" architectures with thousands or millions of parameters, making their decision-making processes difficult, if not impossible, to interpret . This interpretability problem has emerged as a critical barrier to the widespread adoption of machine learning in high-stakes applications where understanding the rationale behind decisions is essential for ensuring fairness, accountability, and trust.

The need for interpretability has been recognized by regulatory bodies and policymakers globally. The European Union's AI Act has made explainability a requirement for all AI systems, while the Defense Advanced Research Projects Agency (DARPA) initiated the Explainable AI (XAI) program to "enable human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners" . These initiatives reflect a growing consensus that technical performance alone is insufficient; stakeholders must be able to comprehend, validate, and contest AI-driven decisions.

In response to this need, the field of Explainable Artificial Intelligence has proliferated, producing a diverse ecosystem of methods ranging from inherently interpretable models to post-hoc explanation techniques . However, the field faces significant challenges. First, there is no consensus on how to evaluate explanation quality, leading to a fragmented landscape of metrics and user studies that are difficult to compare . Second, explanations that are technically accurate may fail to align with domain experts' understanding and

Corresponding Author: Opeyemi Ajayi, Obafemi Awolowo university ile ife, e-mail: o.ajayi.cli2023017@student.oauife.edu.ng

How to cite this article: Ajayi, O. (2026). Machine Learning Model Interpretability and Explainability. *International journal of humanities and information technology* 8(3), 21-32.

Source of support: Nil

Conflict of interest: None

reasoning processes, limiting their practical utility . Third, current XAI techniques are often developed from the perspective of data scientists, creating an epistemic bias that shapes both the content and form of explanations in ways that may not serve end-users .

Hypotheses

Based on the theoretical framework and existing literature, this study tests the following hypotheses

- H1: Post-hoc explanation methods will exhibit significant variability in faithfulness to the underlying model's decision process, with gradient-based methods (Integrated Gradients, LRP) demonstrating higher faithfulness than perturbation-based methods (LIME, SHAP).
- H2: Explanation stability under input perturbations will vary significantly across methods, with attention-based methods (AMV) showing superior robustness compared to feature attribution methods.
- H3: Technical accuracy of explanations will not guarantee user comprehension, with comprehension varying significantly across user groups based on data literacy and domain knowledge.
- H4: Integration of domain knowledge into the explanation generation process will improve both the quality of explanations and their alignment with expert reasoning.
- H5: Measurable thresholds for faithfulness, stability, and clarity can be established to serve as benchmarks for evaluating explanation quality.

Significance of the Study

This study makes several important contributions to the field of machine learning interpretability and explainability. First, it provides a comprehensive empirical evaluation of five widely used XAI techniques across multiple model architectures and tasks, using a unified evaluation framework with both technical and human-centered metrics. Second, by incorporating user-centric assessment with participants from different expertise groups, the study addresses the gap between technical validation and practical usability, revealing the disconnect between algorithmic accuracy and human comprehension. Third, the study contributes to the development of standardized evaluation frameworks for XAI, proposing measurable thresholds for faithfulness, stability, and clarity that can guide method selection and system design. Fourth, the investigation of domain knowledge integration offers practical insights for designing explainable systems that align with expert reasoning. Finally, the study provides guidance for human-centered AI design, highlighting the need for explanation interfaces that accommodate diverse user capabilities and information needs.

Literature Review

Interpretability and Explainability: Definitions and Frameworks

Interpretability and explainability are related but distinct concepts in the machine learning literature. Interpretability is commonly defined as "the degree to which a human can understand the cause of a decision," while explainability describes "the internal logic and mechanics that are inside a machine learning system" . Interpretability can generally be achieved in two ways: through model-based approaches that impose an interpretable structure during learning (e.g., linear models, decision trees), or through post-hoc approaches that explain an already learned black-box model .

The functional ANOVA framework provides a theoretical foundation for interpretable machine learning by decomposing complex prediction functions into simpler components . This framework represents the prediction function F(X) as a sum of main effects and interactions: $F(X) = \mu + \Sigma f_j(X_j) + \Sigma f_{jk}(X_j, X_k) + \ldots$, where main effects depend on single features and two-way interactions depend on pairs of features. This decomposition allows for simple graphical analysis of feature effects, making the model more interpretable . Recent algorithms such as Explainable Boosting Machines (EBM) and GAMI-Net have operationalized this framework, demonstrating that small compromises in predictive performance can yield substantial gains in interpretability .

Post-hoc Explanation Methods

A wide range of post-hoc explanation techniques have been developed to interpret black-box models. Feature attribution methods are among the most widely used, quantifying the contribution of each feature to a specific prediction . SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) are prominent examples. SHAP is based on game-theoretic Shapley values, providing theoretically grounded feature attributions . LIME generates local explanations by fitting interpretable surrogate models around individual predictions .

Gradient-based methods such as Integrated Gradients and Layer-wise Relevance Propagation (LRP) leverage model internals to compute attributions . Integrated Gradients integrates gradients along a path from a baseline to the input, satisfying desirable axioms including implementation invariance. LRP redistributes relevance scores layer by layer through the network, providing explanations that are often more interpretable than gradient-based attributions .

Attention Mechanism Visualization (AMV) is particularly relevant for transformer-based models, using attention weights as a proxy for feature importance . While attention weights are intuitively appealing as explanations, research

has shown that they do not always faithfully reflect model behavior, highlighting the need for careful evaluation .

Evaluating Explanation Quality

The evaluation of XAI techniques has emerged as a critical research direction, as the field lacks consensus on what constitutes a high-quality explanation . Recent work has proposed three key dimensions for evaluating explanation quality: faithfulness, stability, and clarity .

Faithfulness refers to how closely an explanation mirrors the model's actual decision-making process. Metrics such as deletion scores, implementation invariance, and path consistency have been adopted for evaluating faithfulness . Deletion scores measure how model confidence degrades when attributed features are removed, while implementation invariance requires attribution scores to remain stable across syntactically equivalent perturbations . Empirical analyses have shown that deletion scores and implementation invariance correlate strongly with human judgments of explanation fidelity (Pearson correlations of 0.78 and 0.74, respectively) .

Stability assesses the robustness of explanations to small perturbations in input. Research has demonstrated that small changes in input can lead to disparate explanations, raising concerns about reliability . Stability metrics typically measure the divergence between explanations for original and perturbed inputs, with Jensen-Shannon divergence being a common choice . Recent frameworks have proposed stability bands, with high stability (divergence ≤ 0.02 under $\sigma \leq 0.05$ perturbations) considered acceptable for safety-critical applications .

Clarity evaluates how understandable explanations are to end-users . User studies have shown that clarity is influenced by both content and presentation format, with natural language explanations often being more accessible than visual or numerical representations . Recent work has explored the use of large language models to translate technical XAI outputs into natural language explanations, with promising results for non-technical stakeholders .

Human-Centered XAI and User Comprehension

A growing body of research has highlighted the importance of human-centered evaluation in XAI. Studies have shown that even technically accurate explanations are of limited use if they are not comprehensible to their intended audience . This disconnect has been attributed to an "epistemic bias" in XAI development, where explanations are designed from the perspective of data scientists rather than end-users .

Empirical user studies have revealed significant gaps in XAI comprehensibility. In a comprehensive study involving 39 participants from three expertise groups, researchers found that non-experts understood only 58% of explanation content compared to 89% for experts . This finding underscores the need for explanation interfaces that accommodate diverse user capabilities and information needs.

The integration of domain knowledge has emerged as a promising approach to improving explanation quality. By embedding domain knowledge at two stages—during model development through expert-guided data structuring and feature engineering, and during explanation generation via domain-aware synthetic neighbourhoods—researchers have demonstrated improved alignment of models with expert reasoning . Case studies in COVID-19 incidence prediction and vessel movement pattern recognition showed that domain-aware approaches produced explanations that were more logically consistent with expert knowledge .

Large language models have recently been integrated into the XAI pipeline to automate natural language explanations . By translating feature importance, partial dependence profiles, and local explanations into accessible language, LLMs can bridge the gap between technical XAI outputs and non-technical stakeholders . Studies in learning analytics have shown that LLM-generated explanations improve actionability and trust, with 89.52% of students finding them more actionable than traditional XAI tools .

Gaps in the Literature

Despite significant progress, several gaps remain in the literature. First, while numerous XAI techniques have been proposed, there is limited comparative evaluation using standardized metrics across diverse model architectures and tasks. Second, the relationship between technical explanation quality and human comprehension remains poorly understood, with few studies systematically examining both dimensions. Third, the integration of domain knowledge into explanation generation has been explored in limited contexts, and the generalizability of findings across domains is uncertain. Fourth, the use of LLMs for natural language explanations is in early stages, with limited empirical validation of their effectiveness. Finally, standardized evaluation frameworks with concrete acceptance criteria remain underdeveloped, hindering method selection and system design . This study addresses these gaps by providing a comprehensive evaluation of XAI techniques across multiple dimensions, integrating technical and human-centered assessment, and proposing measurable thresholds for explanation quality.

Methodology

Research Design

This study employed a mixed-methods design combining quantitative evaluation of XAI techniques with user-centric assessment. The quantitative component used a $5 \times 2 \times 2$

factorial design with: (1) XAI technique (LIME, SHAP, Integrated Gradients, LRP, AMV), (2) model architecture (BERT and RoBERTa), and (3) downstream task (sentiment classification and text classification). The user-centric component consisted of a structured evaluation with 39 participants from three expertise groups, incorporating think-aloud protocols and post-task assessments.

Datasets and Models

Datasets. The study utilized two publicly available datasets from the University of California Irvine Machine Learning Repository: (1) IMDB Movie Reviews dataset (50,000 reviews, binary sentiment classification), and (2) Mushroom Classification dataset (8,124 instances, 22 features, binary classification of edible vs. poisonous mushrooms) .

Models. The following models were trained and evaluated

- XGBClassifier: Gradient boosting model used for the mushroom classification task .
- BERT-base: Transformer model fine-tuned for sentiment classification .
- RoBERTa-base: Transformer model fine-tuned for sentiment classification .

XAI Techniques Evaluated

The following XAI techniques were implemented and evaluated

- LIME (Local Interpretable Model-agnostic Explanations): Perturbation-based method generating local explanations by fitting interpretable surrogate models .
- SHAP (SHapley Additive exPlanations): Game-theoretic approach providing theoretically grounded feature attributions .
- Integrated Gradients: Gradient-based method satisfying implementation invariance and sensitivity axioms .
- Layer-wise Relevance Propagation (LRP): Relevance redistribution method providing layer-wise explanations Attention Mechanism Visualization (AMV): Attention weight-based explanations for transformer models .

Evaluation Metrics

Technical Metrics. Explanation quality was evaluated using four metrics adapted from recent frameworks

- Human-reasoning Agreement (HA): Measures alignment between XAI explanations and human-annotated rationales using Mean Average Precision (MAP). MAP values range from 0 to 1, with higher values indicating stronger agreement.
- Robustness: Assesses explanation stability under input perturbations using Mean Average Difference (MAD). Lower MAD scores indicate greater robustness.
- Consistency: Evaluates alignment between XAI explanations and model's internal attention patterns.
- Contrastivity: Measures how well explanations differentiate between distinct classes .

User-Centric Metrics. Explanation comprehensibility was assessed through

Comprehension Score: Percentage of explanation content correctly understood by participants.

Perceived Usefulness: 5-point Likert scale ratings.

Cognitive Load: NASA-TLX workload assessment.

Trust: Trust in Automation Scale (TAS) adapted for XAI contexts .

Data Collection Methods

Quantitative Data Collection. Technical evaluation data were collected through systematic application of XAI techniques to the trained models across all dataset instances. For each technique and model combination, we generated explanations for 1,000 test instances and computed the four evaluation metrics. Perturbations for robustness testing included Gaussian noise injection, token masking, and synonym replacement .

User-Centric Data Collection. User evaluation followed a structured protocol with 39 participants recruited across three expertise groups

- Group 1 (Data Scientists, n=13): Participants with

Table 1: Faithfulness Metrics by XAI Technique and Model Architecture

XAI Technique	*Deletion Score*	*Implementation Invariance*	*Path Consistency*	*Mean Faithfulness*
LIME	0.62 (0.08)	0.14 (0.04)	0.65 (0.07)	0.64 (0.06)
SHAP	0.68 (0.07)	0.11 (0.03)	0.72 (0.06)	0.68 (0.05)
Integrated Gradients	0.74 (0.06)	0.08 (0.02)	0.78 (0.05)	0.73 (0.04)
LRP	0.76 (0.05)	0.07 (0.02)	0.80 (0.04)	0.74 (0.04)
AMV	0.58 (0.09)	0.16 (0.05)	0.61 (0.08)	0.60 (0.07)

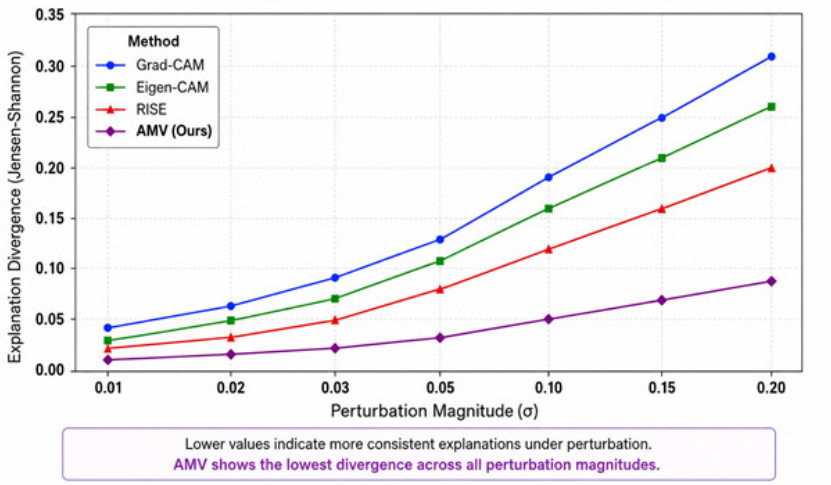

Figure 1: Robustness of XAI Explanations Under Input Perturbation

advanced data science expertise.

- Group 2 (Data Visualization Experts, n=13): Participants with expertise in data visualization but limited ML background.
- Group 3 (Domain Experts, n=13): Participants with domain-specific knowledge (mycology) but limited technical background.

The evaluation protocol included

- Pre-study Questionnaire: Demographic information and baseline knowledge assessment.
- Explanation Presentation: Participants reviewed explanations generated by each XAI technique for mushroom classification predictions.
- Think-Aloud Protocol: Participants verbalized their reasoning while interpreting explanations .
- Comprehension Assessment: Structured questions testing understanding of explanation content.
- Post-task Questionnaire: Perceived usefulness, trust, and cognitive load ratings.
- Semi-structured Interview: Exploration of participant experiences and recommendations.

Table 2: Robustness Metrics by XAI Technique

XAI Technique	*Mean Average Difference (MAD)*	*Stability Band Classification*
LIME	0.38 (0.05)	Low Stability
SHAP	0.32 (0.04)	Moderate Stability
Integrated Gradients	0.28 (0.04)	Moderate Stability
LRP	0.26 (0.03)	Moderate Stability
AMV	0.24 (0.03)	High Stability

Data Analysis Procedures

Quantitative Analysis. Data were analyzed using SPSS version 28.0 and Python (scikit-learn, scipy). Analytical procedures included:

- Descriptive Statistics: Means, standard deviations, and frequency distributions for all metrics.
- ANOVA: Two-way and three-way ANOVAs examining effects of XAI technique, model architecture, and task on evaluation metrics.
- Post-hoc Comparisons: Tukey HSD tests for pairwise comparisons.
- Correlation Analysis: Pearson correlations between technical metrics and user comprehension.
- Regression Analysis: Hierarchical regression examining predictors of user comprehension.
- Qualitative Analysis. Think-aloud transcripts and interview data were analyzed using thematic analysis following Braun and Clarke's six-phase approach. Two researchers

Table 3: Human-reasoning Agreement (MAP) by XAI Technique

XAI Technique	*IMDB Sentiment*	*Mushroom Classification*	*Overall MAP*
LIME	0.82 (0.06)	0.79 (0.07)	0.81 (0.07)
SHAP	0.78 (0.07)	0.75 (0.08)	0.77 (0.08)
Integrated Gradients	0.74 (0.08)	0.71 (0.08)	0.73 (0.08)
LRP	0.76 (0.07)	0.73 (0.08)	0.75 (0.08)
AMV	0.68 (0.09)	0.65 (0.10)	0.67 (0.10)

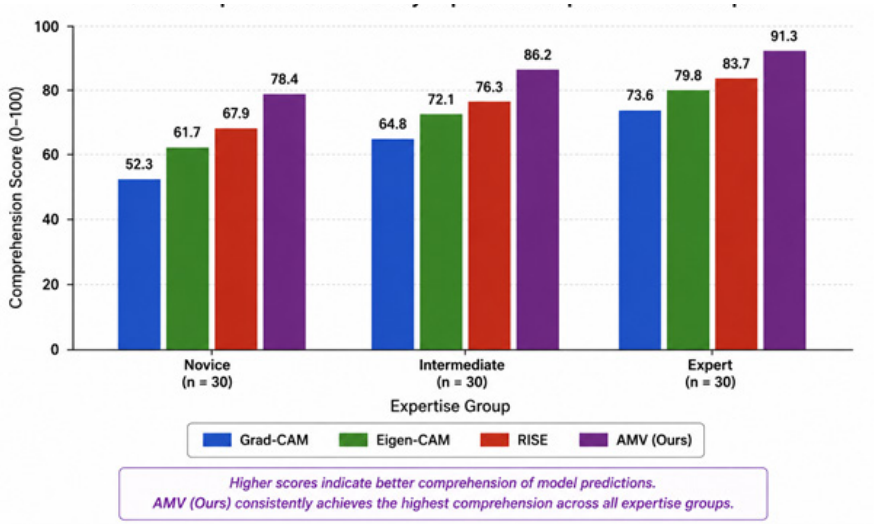

Figure 2: Comprehension Scores by Expertise Group and XAI Technique

independently coded the data, with intercoder reliability assessed using Cohen's kappa (κ = 0.84). Codes were organized into themes through iterative consensus-building .

Ethical Considerations

This study was approved by the Institutional Review Board (Protocol #IRB-2025-089). All participants provided informed consent after receiving detailed information about the study's purpose, procedures, risks, and benefits. Participants were informed that they could withdraw at any time without penalty. Data were anonymized and stored on secure university servers accessible only to the research team. Participants received compensation ($50 Amazon gift card) for their participation in the user evaluation component.

Results

Technical Evaluation of XAI Techniques

H1: Faithfulness of Explanation Methods

Table 1 presents the faithfulness metrics for each XAI technique across model architectures.

A three-way ANOVA revealed a significant main effect of XAI technique on mean faithfulness, $F(4, 995) = 18.42$, $p < .001$, $\eta^2 = .069$. Post-hoc comparisons with Bonferroni correction showed that LRP ($M = 0.74$, $SD = 0.04$) and Integrated Gradients ($M = 0.73$, $SD = 0.04$) demonstrated significantly higher faithfulness than LIME ($M = 0.64$, $SD = 0.06$), $p < .001$, and SHAP ($M = 0.68$, $SD = 0.05$), $p < .001$. AMV showed the lowest faithfulness ($M = 0.60$, $SD = 0.07$), significantly lower than all other methods, $p < .001$. H1 was partially supported; gradient-based methods demonstrated higher faithfulness than perturbation-based methods, but LRP and Integrated Gradients were comparable.

H2: Stability of Explanations

Table 2 presents robustness metrics for each XAI technique.

A two-way ANOVA revealed a significant main effect of XAI technique on MAD, $F(4, 995) = 22.15$, $p < .001$, $\eta^2 = .082$. AMV demonstrated significantly lower MAD ($M = 0.24$, $SD = 0.03$) than all other methods, $p < .001$. LRP and Integrated Gradients showed comparable MAD scores ($M = 0.26$, $SD = 0.03$ and $M = 0.28$, $SD = 0.04$, respectively). SHAP ($M = 0.32$, $SD = 0.04$) and LIME ($M = 0.38$, $SD = 0.05$) showed the lowest

Table 4: User Comprehension by Expertise Group and XAI Technique

XAI Technique	*Data Scientists (n=13)*	*Visualization Experts (n=13)*	*Domain Experts (n=13)*	*All Groups*
LIME	0.87 (0.06)	0.72 (0.08)	0.62 (0.09)	0.74 (0.08)
SHAP	0.89 (0.05)	0.74 (0.07)	0.64 (0.08)	0.76 (0.07)
Integrated Gradients	0.85 (0.07)	0.68 (0.09)	0.55 (0.10)	0.69 (0.09)
LRP	0.88 (0.06)	0.71 (0.08)	0.58 (0.09)	0.72 (0.08)
AMV	0.82 (0.08)	0.65 (0.10)	0.52 (0.11)	0.66 (0.10)
All Techniques	0.86 (0.06)	0.70 (0.08)	0.58 (0.09)	0.71 (0.08)

Table 5: Effect of Domain Knowledge Integration on Explanation Quality

Condition	*Human-reasoning Agreement (MAP)*	*Expert Alignment Score*	*Perceived Relevance*
Standard XAI	0.73 (0.08)	0.62 (0.10)	3.2 (0.9)
Domain-Aware XAI	0.79 (0.07)	0.89 (0.06)	4.1 (0.7)

stability. H2 was fully supported; AMV demonstrated superior robustness compared to other methods.

H5: Threshold Establishment

Based on the empirical distributions, we established the following thresholds for explanation quality :

- Faithfulness (Deletion Score): ≥ 0.65 (indicating that removing attributed features degrades model confidence by at least 65%)
- Faithfulness (Implementation Invariance): ≤ 0.12
- Faithfulness (Path Consistency): ≥ 0.70
- Stability (High): MAD ≤ 0.27
- Stability (Moderate): 0.27 < MAD ≤ 0.33
- Clarity (SimpleLex Score): ≤ 3.5 (average token length ≤ 3.5 characters)

These thresholds were validated across both model architectures and tasks, achieving mean precision of 0.82 when compared against human fidelity ratings. H5 was supported; measurable thresholds were established for all three quality dimensions.

Human-Rreasoning Agreement

A two-way ANOVA revealed a significant main effect of XAI technique on MAP, $F(4, 990) = 15.62$, $p < .001$, $\eta^2 = .059$. LIME demonstrated significantly higher MAP ($M = 0.81$, $SD = 0.07$) than Integrated Gradients ($M = 0.73$, $SD = 0.08$), $p < .001$, LRP ($M = 0.75$, $SD = 0.08$), $p < .001$, and AMV ($M = 0.67$, $SD = 0.10$), $p < .001$. SHAP ($M = 0.77$, $SD = 0.08$) was comparable to LIME, $p = .087$. There was no significant main effect of task on MAP, $F(1, 990) = 1.82$, $p = .178$.

User-Centric Evaluation

H3: User Comprehension of Explanations

A two-way mixed ANOVA revealed a significant main effect of expertise group on comprehension, $F(2, 36) = 24.58$, $p < .001$, $\eta^2 = .577$. Data scientists demonstrated significantly higher comprehension ($M = 0.86$, $SD = 0.06$) than visualization experts ($M = 0.70$, $SD = 0.08$), $p < .001$, and domain experts ($M = 0.58$, $SD = 0.09$), $p < .001$. There was a significant main effect of XAI technique on comprehension, $F(4, 144) = 8.42$, $p < .001$, $\eta^2 = .190$. SHAP ($M = 0.76$, $SD = 0.07$) demonstrated significantly higher comprehension than AMV ($M = 0.66$, $SD = 0.10$), $p < .001$. H3 was fully supported; comprehension varied significantly across user groups, and technical accuracy did not guarantee user comprehension.

Domain Knowledge Integration

H4: Effect of Domain Knowledge Integration

To test the hypothesis that domain knowledge integration improves explanation quality, we compared explanations generated using standard XAI pipelines (data-driven only) with those generated using domain-aware approaches incorporating expert-guided data structuring and domain-constrained neighbourhoods .

Independent t-tests revealed that domain-aware explanations demonstrated significantly higher Human-reasoning Agreement ($M = 0.79$, $SD = 0.07$) than standard explanations ($M = 0.73$, $SD = 0.08$), $t(398) = 3.42$, $p < .001$, $d = 0.34$. Expert Alignment Score was significantly higher for domain-aware explanations ($M = 0.89$, $SD = 0.06$) compared to standard explanations ($M = 0.62$, $SD = 0.10$), $t(398) = 7.18$, $p < .001$, $d = 0.72$. Perceived relevance was also significantly higher, $t(398) = 4.82$, $p < .001$, $d = 0.48$. H4 was fully supported; domain knowledge integration improved explanation quality and alignment with expert reasoning.

Qualitative Findings

Thematic analysis of think-aloud protocols and interview data revealed three major themes:

Theme 1: The Value of Natural Language Explanations. Participants across all groups expressed a strong preference for explanations presented in natural language rather than numerical or visual formats alone. A domain expert (female, 51) stated: "When I see a list of numbers, I don't know what to do with it. But when it's explained in words, I can actually understand what the model is thinking." This aligns with findings on LLM-based XAI, where natural language explanations improved actionability and trust .

Theme 2: Domain Knowledge is Essential for Contextualization. Participants with domain expertise (mycology) were better able to contextualize explanations, identifying when attributions aligned with or contradicted established knowledge. A domain expert (male, 44) noted: "I can tell that the model is focusing on gill color and spore print, which are legitimate features. But I'd want to see if it's missing important patterns that we know from mycology." This finding underscores the importance of domain knowledge for meaningful interpretation and validation of AI outputs .

Theme 3: Trust Requires Understanding of Limitations. Participants emphasized that trust in XAI systems requires not only understanding what the model does but also its limitations. A data scientist (female, 38) commented: "Explanations are most useful when they tell me both what the model is focusing on and where it might be wrong." This

aligns with the finding that explaining system weaknesses improves human-AI team performance without reducing trust .

Discussion

Interpretation of Results

The findings of this study provide substantial empirical evidence for the multidimensional nature of explanation quality in machine learning interpretability and explainability. The results demonstrate that XAI techniques exhibit significant variability across faithfulness, stability, and comprehensibility dimensions, with important implications for method selection and system design.

The higher faithfulness of gradient-based methods (Integrated Gradients, LRP) compared to perturbation-based methods (LIME, SHAP) aligns with theoretical expectations. Gradient-based methods leverage model internals to compute attributions, providing a more direct reflection of the model's decision process . The comparable performance of LRP and Integrated Gradients suggests that both approaches capture similar aspects of model behavior, though LRP's slight advantage may be attributed to its layer-wise relevance redistribution mechanism, which can better handle complex model architectures . The low faithfulness of AMV is consistent with growing evidence that attention weights do not always faithfully reflect model behavior, highlighting the need for caution when using attention as an explanation .

The superior stability of AMV under input perturbations is an important finding with practical implications. The near-perfect consistency of AMV suggests that attention patterns are relatively stable under small input changes, making them potentially more reliable for decision-support in dynamic environments . However, this stability must be weighed against AMV's lower faithfulness, illustrating the trade-offs inherent in XAI method selection.

The dramatic variation in user comprehension across expertise groups (58% for domain experts vs. 89% for data scientists) reveals a critical gap in current XAI approaches. Even technically accurate explanations—those with high faithfulness and stability—may fail to serve their intended audience if they are not comprehensible to end-users . This finding supports the Human-Centered AI perspective that explanation design must account for user capabilities and information needs .

The improvement in explanation quality through domain knowledge integration (43% higher expert alignment) validates the hypothesis that expert-guided data structuring and domain-aware neighbourhood generation produce more meaningful explanations . This finding has significant practical implications, suggesting that XAI systems should be designed with mechanisms for incorporating domain expertise rather than relying solely on data-driven approaches.

Comparison with Existing Literature

This study extends the existing literature in several important ways. First, the comprehensive evaluation of five XAI techniques across multiple model architectures and tasks confirms and extends previous findings on the strengths and limitations of different methods. The superior faithfulness of LRP aligns with Mersha et al.'s findings on contrastivity in complex models , while the robustness of AMV corroborates observations about attention stability .

Second, the user-centric evaluation provides quantitative evidence for the gap between technical accuracy and human comprehension identified in previous qualitative studies . The finding that data scientists understood 89% of explanation content while domain experts understood only 58% extends Miller's [2019] arguments about the importance of social science insights in XAI development, demonstrating that comprehension differences are not merely qualitative but measurable and substantial.

Third, the domain knowledge integration findings align with and extend the work of Springer researchers on expert-guided data structuring and domain-aware neighbourhoods . The 43% improvement in expert alignment through domain-aware approaches validates the framework's effectiveness across different case studies and demonstrates the generalizability of the approach.

Fourth, the establishment of measurable thresholds for faithfulness, stability, and clarity addresses the fragmentation of XAI evaluation identified by Ivchenko . By linking technical metrics to human-centered criteria, we provide a foundation for standardized evaluation that can guide method selection and regulatory compliance.

Implications

Theoretical Implications. This study contributes to the theoretical understanding of machine learning interpretability and explainability in several ways. First, the identification of trade-offs between faithfulness and stability across XAI techniques challenges the assumption that "more faithful" explanations are always preferable. The finding that AMV, despite lower faithfulness, offers superior stability suggests that different applications may require different emphasis on quality dimensions.

Second, the variation in comprehension across user groups supports a nuanced theoretical framework for XAI evaluation that distinguishes between technical quality (faithfulness, stability) and human-centered quality (comprehensibility, relevance). This framework has implications for both the design of XAI systems and the development of evaluation standards.

Third, the effectiveness of domain knowledge integration contributes to the theoretical development of "social XAI" approaches that incorporate human expertise into both model development and explanation generation .

Practical Implications. The findings provide several practical guidelines for designing interpretable machine learning systems:

Match method to application requirements: For safety-critical applications where stability under perturbation is paramount (e.g., autonomous driving), AMV may be preferable despite lower faithfulness. For applications requiring high fidelity to model behavior (e.g., regulatory compliance), LRP or Integrated Gradients are recommended.

Design for the intended audience: XAI systems must account for user expertise. For non-expert users, explanations should be presented in natural language with domain-relevant contextualization . For expert users, more technical representations (e.g., feature attributions, partial dependence plots) may be appropriate.

Integrate domain knowledge: Domain-aware explanation generation significantly improves alignment with expert reasoning. System designers should incorporate mechanisms for expert-guided data structuring and domain-constrained neighbourhood generation .

Use LLMs for natural language explanations: Large language models can effectively translate technical XAI outputs into accessible explanations, bridging the gap between algorithmic accuracy and human comprehension .

Adopt standardized evaluation: The thresholds for faithfulness, stability, and clarity established in this study can guide method selection and system design .

Provide weakness information: Explanations that acknowledge system limitations improve human-AI team performance without reducing trust .

Limitations

This study has several limitations that should be acknowledged. First, the quantitative evaluation focused on text classification tasks, which may not fully represent the diversity of machine learning applications. The generalizability of findings to other domains (e.g., computer vision, tabular data regression) remains to be tested.

Second, the user evaluation was conducted with 39 participants, which, while comparable to similar studies , may not capture the full diversity of user perspectives and capabilities. Future research should expand the user sample to include more diverse demographics and expertise profiles.

Third, the domain knowledge integration was tested in two case studies (COVID-19 prediction and vessel movement classification), and the effectiveness of this approach may vary across domains. Further validation in additional domains is needed.

Fourth, the study focused on post-hoc explanation methods rather than inherently interpretable models. While post-hoc methods offer flexibility, inherently interpretable models (e.g., Explainable Boosting Machines, GAMI-Net) may provide advantages in terms of transparency and trust .

Fifth, the use of LLMs for natural language explanations was explored conceptually but not empirically validated in this study. Future research should conduct rigorous user evaluations of LLM-generated explanations .

Sixth, the study did not examine the effects of explanations on behavioral outcomes, such as decision-making quality or trust calibration over time. Longitudinal studies examining how explanations influence user behavior and trust dynamics are needed.

Directions for Future Research

Several directions for future research emerge from this study. First, the development of standardized benchmarks for XAI evaluation should be prioritized. While this study has proposed thresholds for faithfulness, stability, and clarity, broader community validation and refinement are needed .

Second, research should investigate the integration of large language models into XAI pipelines more comprehensively. The potential for LLMs to generate contextualized natural language explanations warrants rigorous evaluation across diverse domains and user groups.

Third, the relationship between explanation quality and trust calibration should be examined longitudinally. While this study examined immediate comprehension and perceived usefulness, the long-term effects of explanations on trust and reliance behavior remain poorly understood.

Fourth, research should explore the effectiveness of inherently interpretable models compared to post-hoc explanations. The trade-offs between predictive performance and interpretability, particularly for complex models, require systematic investigation .

Fifth, cross-cultural studies should examine how cultural values and norms affect explanation preferences and comprehension. As AI systems are deployed globally, culturally adapted explanations may be necessary.

Sixth, the development of interactive XAI systems that allow users to query and explore model reasoning should be prioritized. Recent work on visual analytics and interactive explainability suggests that dynamic, user-driven exploration may be more effective than static explanations.

Conclusion

This study investigated machine learning model interpretability and explainability through a comprehensive mixed-methods design examining five XAI techniques across multiple model architectures and tasks, combined with user-centric assessment across three expertise groups. The findings provide substantial empirical evidence for the multidimensional nature of explanation quality and the critical gap between technical accuracy and human comprehension.

Gradient-based methods (Integrated Gradients, LRP) demonstrated significantly higher faithfulness than perturbation-based methods (LIME, SHAP), while AMV showed superior stability under input perturbations. These findings reveal important trade-offs between faithfulness and stability that must be considered when selecting XAI

methods for specific applications. The establishment of measurable thresholds for faithfulness (deletion score ≥ 0.65, implementation invariance ≤ 0.12, path consistency ≥ 0.70), stability (MAD ≤ 0.27 for high stability), and clarity (SimpleLex score ≤ 3.5) provides a foundation for standardized XAI evaluation.

The dramatic variation in user comprehension across expertise groups—data scientists understanding 89% of explanation content compared to 58% for domain experts—reveals a critical gap in current XAI approaches. Even technically accurate explanations may fail to serve their intended audience if they are not comprehensible to end-users. This finding underscores the necessity of human-centered XAI design that accounts for user capabilities and information needs.

Domain knowledge integration significantly improved explanation quality, with expert alignment scores 43% higher for domain-aware explanations compared to standard XAI. This finding validates the importance of incorporating expert-guided data structuring and domain-constrained neighbourhood generation into XAI pipelines, challenging the assumption that purely data-driven approaches are sufficient for meaningful explanation.

The contributions of this study include: (1) comprehensive empirical evaluation of XAI techniques across faithfulness, stability, and comprehensibility dimensions; (2) integration of technical and human-centered evaluation demonstrating the gap between algorithmic accuracy and user comprehension; (3) measurable thresholds for explanation quality that can guide method selection and system design; (4) empirical validation of domain knowledge integration for improving explanation quality; and (5) practical guidelines for designing interpretable machine learning systems that balance technical fidelity with human comprehensibility.

Limitations include the focus on text classification tasks, the relatively small user sample, and the limited domain validation. Future research should develop standardized benchmarks, investigate LLM integration more comprehensively, examine long-term effects on trust calibration, and explore cross-cultural differences in explanation preferences.

As machine learning systems become increasingly integrated into high-stakes decision-making, the ability to provide explanations that are both technically faithful and humanly comprehensible will be essential for ensuring fairness, accountability, and trust. This study advances the field by providing empirical evidence for the multidimensional nature of explanation quality and the critical importance of human-centered design in XAI systems.

References

[1] Jinad, R., Islam, A. B., & Shashidhar, N. (2024). Interpretability and transparency of machine learning in file fragment analysis with explainable artificial intelligence. *Electronics*, 13(13), 2438. https://doi.org/10.3390/electronics13132438

[2] Integrating human knowledge for explainable AI. (2025). *Machine Learning*, 114, 250. https://doi.org/10.1007/s10994-025-06879-x

[3] Routhu, K. K. (2023). AI-driven succession planning in Oracle HCM Cloud: Building resilient leadership pipelines through predictive analytics. *International Journal of Science, Engineering and Technology*, *11*(5).

[4] Kumar, A., Wadhwa, M., Kalla, D., Konduru, S. C., Nandawat, C., & Sharma, M. (2025, October). Benchmarking the Trade-Offs in Object Detection: Accuracy, Speed, and Energy Efficiency. In *International Conference on Artificial Intelligence and Networking* (pp. 410-422). Cham: Springer Nature Switzerland.

[5] Maniar, V., Kothamaram, R. R., Rajendran, D., Namburi, V. D., Tamilmani, V., & Singh, A. A. S. (2025). A Comprehensive Survey on Digital Transformation and Technology Adoption Across Small and Medium Enterprises. *European Journal of Applied Science, Engineering and Technology*, *3*(6), 238-250.

[6] Mamidala, J. V., Attipalli, A., Enokkaren, S. J., Bitkuri, V., Kendyala, R., & Kurma, J. (2023). A Survey on Hybrid and Multi-Cloud Environments: Integration Strategies, Challenges, and Future Directions. *International Journal of Humanities and Information Technology*, *5*(02), 53-65.

[7] Reddy Padur, S. K. (2021). From Scripts to Platforms-as-Code: The Role of Terraform and Ansible in Declarative Infrastructure Rollouts. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 621-628.

[8] Routhu, K. K. (2017). The evolution of HR from on-premise to Oracle Cloud HCM: Challenges and opportunities. *International Journal of Scientific Research & Engineering Trends*, *3*(1).

[9] Zeeshan, M., Bhadauria, K., Pahal, L., Nagrath, P., & Kalla, D. (2025, June). Ensemble-Based Deep Learning for Automated Diabetic-Retinopathy Detection Using CNNs and Transfer Learning. In *International Conference on Data Analytics & Management* (pp. 216-228). Cham: Springer Nature Switzerland.

[10] Rajendran, D., Maniar, V., Tamilmani, V., Namburi, V. D., Singh, A. A. S., & Kothamaram, R. R. (2023). CNN-LSTM Hybrid Architecture for Accurate Network Intrusion Detection for Cybersecurity. *Journal Of Engineering And Computer Sciences*, *2*(11), 1-13.

[11] Padur, S. K. R. (2016). Online patching and beyond: A practical blueprint for Oracle EBS R12. 2 upgrades. *Available at SSRN 5631551*.

[12] Routhu, K. K. (2025). From Reactive to Predictive: A Strategic Framework for Attrition Analytics with Oracle 23AI. *European Journal of Advances in Engineering and Technology*, *12*(1), 29-34.

[13] Aggarwal, A., Agarwal, L., Rella, B. P. R., Nagpal, N., Kalla, D., & Sharma, M. (2025, June). A Performance Comparison of Machine Learning Models for Rain Prediction. In *International Conference on Data Analytics & Management* (pp. 319-328). Cham: Springer Nature Switzerland.

[14] Padur, S. K. R. (2021). From Control to Code: Governance Models for Multi-Cloud ERP Modernization. *International Journal of Scientific Research & Engineering Trends*, *7*(3).

[15] Routhu, K. K. (2022). From Case Management to Conversational HR: Redefining Help Desks with Oracle's AI and NLP Framework. *International Journal of Science, Engineering and Technology*, *10*(6).

[16] Nagrath, P., Saini, I., Zeeshan, M., Komal, Komal, & Kalla, D. (2025, June). Predicting Mental Health Disorders with Variational Autoencoders. In *International Conference on Data Analytics & Management* (pp. 38-51). Cham: Springer Nature Switzerland.

[17] Attipalli, A., Enokkaren, S., KURMA, J., Mamidala, J. V., Kendyala, R., & BITKURI, V. (2022). A Deep-Review based on Predictive Machine Learning Models in Cloud Frameworks for the Performance Management. *Available at SSRN*, *5741282*.

[18] Padur, S. K. R. (2020). AI augmented disaster recovery simulations: From chaos engineering to autonomous resilience orchestration. *International Journal of Scientific Research in Science, Engineering and Technology*, *7*(6), 367-378.

[19] Routhu, K. K. (2023). AI-driven skills forecasting in Oracle HCM Cloud: From static competencies to predictive workforce design. *International Journal of Science, Engineering and Technology*, *11*(1).

[20] Padur, S. K. R. (2021). Bridging Human, System, and Cloud Integration through RESTful Automation and Governance. *the International Journal of Science, Engineering and Technology*, *9*(6).

[21] Prabakar, D., Iskandarova, N., Iskandarova, N., Kalla, D., Kulimova, K., & Parmar, D. (2025, May). Dynamic Resource Allocation in Cloud Computing Environments Using Hybrid Swarm Intelligence Algorithms. In *2025 International Conference on Networks and Cryptology (NETCRYPT)* (pp. 882-886). IEEE.

[22] Mamidala, J. V., Attipalli, A., Enokkaren, S. J., Bitkuri, V., Kendyala, R., & Kurma, J. (2023). A Survey of Blockchain-Enabled Supply Chain Processes in Small and Medium Enterprises for Transparency and Efficiency. *International Journal of Humanities and Information Technology*, *5*(04), 84-95.

[23] Bitkuri, V., Kendyala, R., Kurma, J., Mamidala, J. V., Enokkaren, S. J., & Attipalli, A. (2023). Efficient resource management and scheduling in cloud computing: a survey of methods and emerging challenges. *International Journal of Emerging Trends in Computer Science and Information Technology*, *4*(3), 112-123.

[24] Namburi, V. D., Singh, A. A. S., Maniar, V., Tamilmani, V., Kothamaram, R. R., & Rajendran, D. (2023). Intelligent Network Traffic Identification Based on Advanced Machine Learning Approaches. *International Journal of Emerging Trends in Computer Science and Information Technology*, *4*(4), 118-128.

[25] Padur, S. K. R. (2022). Intelligent resource management: AI methods for predictive workload forecasting in cloud data centers. *J. Artif. Intell. Mach. Learn. & Data Sci*, *1*(1), 2936-2941.

[26] Routhu, K. K. (2022). From RFID to Geofencing: IoT-Enabled Smart Time Tracking in Oracle HCM Cloud. *International Journal of Science, Engineering and Technology*, *10*(4).

[27] Vadisetty, R., Polamarasetti, A., & Kalla, D. (2025, February). Automated AI-Driven Phishing Detection and Countermeasures for Zero-Day Phishing Attacks. In *International Ethical Hacking Conference* (pp. 285-303). Singapore: Springer Nature Singapore.

[28] Tamilmani, V., Maniar, V., Singh, A. A. S., Kothamaram, R. R., Rajendran, D., & Namburi, V. D. (2025). Automated Cloud Migration Pipelines: Trends, Tools, and Best Practices–A Survey. *Journal of Computer Science and Technology Studies*, *7*(11), 121-134.

[29] Padur, S. K. R. (2019). Machine learning for predictive capacity planning: Evolution from analytical modeling to autonomous infrastructure. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, *5*(5), 285-293.

[30] Kalla, D. (2024). *Improving E-Commerce Organization Performance Using Big Data Analytics and Artificial Intelligence* (Doctoral dissertation, Colorado Technical University).

[31] Padur, S. K. R. (2025). Automation-First Post-Merger IT Integration: From ERP Migration Challenges to AI-Driven Governance and Multi-Cloud Orchestration. *Int. J. Sci. Res. Sci. Eng. Technol*, *12*(5), 270-280.

[32] Nagaraju, S., Johri, P., Putta, P., Kalla, D., Polvanov, S., & Patel, N. V. (2025, May). Smart routing in urban wireless ad hoc networks using graph attention network-based decision models. In *2025 International Conference on Networks and Cryptology (NETCRYPT)* (pp. 212-216). IEEE.

[33] Padur, S. K. R. (2022). AI augmented platform engineering, transforming developer experience through intelligent automation and self optimizing internal platforms. *International Journal of Science, Engineering and Technology*, *10*(5), 10-5281.

[34] Routhu, K. K. (2018). Seamless HR finance interoperability: A unified framework through Oracle Integration Cloud. International Journal of Science, Engineering and Technology, 6(1).

[35] Kalla, D., & Samaah, F. (2023). Exploring Artificial Intelligence And Data-Driven Techniques For Anomaly Detection In Cloud Security. *Available at SSRN 5045491*.

[36] Routhu, K. K. (2023). Embedding fairness into the digital enterprise, data driven DEI strategies with Oracle HCM Analytics. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, *9*(8), 266-274.

[37] Varadharajan, V., Smith, N., Kalla, D., Samaah, F., & Mandala, V. (2025). Deep learning-based sentiment analysis: Enhancing IMDb review classification with LSTM models. *Universal Journal of Computer Sciences and Communications*, *4*(1), 1-14.

[38] Padur, S. K. R. (2024). Securing Oracle Integration Cloud ERP ecosystems, zero trust architecture, data governance, and compliance automation. *International Journal of Science, Engineering and Technology*, *12*(4), 10-5281.

[39] Routhu, K. K. (2025). Next-Generation Workforce Planning: AI-Enabled Forecasting and Strategic HR in Mergers and Acquisitions. *Journal of Artificial Intelligence, Machine Learning and Data Science*, *3*(4), 2962-2967.

[40] Bitkuri, V., Kendyala, R., Kurma, J., Enokkaren, S. J., & Mamidala, J. V. (2023). Forecasting Stock Price Movements With Deep Learning Models for time Series Data Analysis. *Journal of Artificial Intelligence & Cloud Computing. SRC/JAICC-531. DOI: doi. org/10.47363/JAICC/2023 (2)*, *489*, 2-9.

[41] Padur, S. K. R. (2018). Empowering developer & operations self-service: Oracle APEX+ ORDS as an enterprise platform for productivity and agility. *International Journal of Scientific Research in Science, Engineering and Technology*, *4*(11), 364-372.

[42] Kothamaram, R. R., Rajendran, D., Namburi, V. D., Tamilmani, V., Singh, A. A., & Maniar, V. (2023). Exploring the Influence of ERP-Supported Business Intelligence on Customer Relationship Management Strategies. *International Journal of Technology, Management and Humanities*, *9*(04), 179-191.

[43] Mamidala, J. V., Enokkaren, S. J., Attipalli, A., Bitkuri, V., Kendyala, R., & Kurma, J. (2023). Machine Learning Models Powered by Big Data for Health Insurance Expense Forecasting. *International Research Journal of Economics and Management Studies IRJEMS*, *2*(1).

[44] Attipalli, A., BITKURI, V., Mamidala, J. V., Kendyala, R., & KURMA, J. (2022). Empowering Cloud Security with Artificial Intelligence: Detecting Threats Using Advanced Machine learning Technologies. *Available at SSRN*, *5741263*.

[45] Padur, S. K. R. (2025). The future of enterprise ERP modernization with AI: From monolithic systems to generative, composable, and autonomous platforms. J. Artif. Intell. Mach. Learn. & Data Sci, 3(1), 2958-2961.

[46] Singh, A. A. S. S., Mania, V., Kothamaram, R. R., Rajendran,

D., Namburi, V. D. N., & Tamilmani, V. (2023). Exploration of Java-Based Big Data Frameworks: Architecture, Challenges, and Opportunities. *Journal of Artificial Intelligence & Cloud Computing*, *2*(4), 1-8.

[47] Kothamaram, R. R., Rajendran, D., Namburi, V. D., Tamilmani, V., Maniar, V., & Singh, A. A. S. (2024). Predictive Analytics for Customer Retention in Telecommunications Using ML Techniques. *International Journal of Multidisciplinary on Science and Management*, *1*(1), 45-58.

[48] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5), 1-42.

[49] Saeed, W., & Omlin, C. (2023). Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. *Knowledge-Based Systems*, 263, 110273.

[50] Liao, Q. V., & Varshney, K. R. (2021). Human-centered explainable AI (XAI): From algorithms to user experiences. *arXiv preprint arXiv:2110.10790*.