American Journal of Advanced Technology and Engineering Solutions

Volume: 6; **Issue:** 1
Pages: 332-368
eISSN: 3067-5146

open access

Crossref

Artificial Intelligence Driven Explainable Machine Learning for High-Stakes Decision Support: SHAP Interpretability and Robustness Testing in Healthcare and Finance

Jarif Ul Alam[1];

[1]. *Doctor of Philosophy (Ph.D.) in Computer Information Systems (Continuing), Louisiana Tech University, USA; Email: jua003@latech.edu*

Doi: 10.63125/mv67zd78

Received: *09 October 2025;* ***Revised:*** *13 November 2025;* ***Accepted:*** *12 December 2025;* ***Published:*** *18 January 2026*

Abstract

This study addressed a persistent challenge in cloud-deployed, enterprise decision-support systems: even when machine-learning models exhibit strong predictive performance, users may under-rely or inconsistently rely on recommendations because post hoc explanations are perceived as unclear or unstable, constraining trust, defensibility, and confident action. The purpose of the study was to examine how explanation quality perceptions, specifically Perceived SHAP Interpretability (PSI) and Perceived Explanation Robustness (PER), are associated with Trust in AI decision support (TRU), Decision Confidence (DCF), and Intention to Rely or Use (INT) in high-stakes, case-based decision scenarios. A quantitative, cross-sectional, case-based design was employed using two enterprise contexts: healthcare clinical risk decision support and financial risk governance decision support. Standardized SHAP explanation artifacts were presented alongside decision vignettes, and a 5-point Likert-scale instrument measured PSI, PER, TRU, DCF, and INT, with controls for professional experience, AI familiarity, and domain group. The final sample consisted of N = 240 screened respondents (52.1% healthcare; 47.9% finance), with a mean professional experience of 7.8 years (SD = 4.6) and moderate AI familiarity (M = 3.62, SD = 0.84). The analytic approach combined descriptive statistics, reliability analysis, Pearson correlations, and a series of hierarchical multiple regression models. Mediation relationships were examined using regression-based mediation logic through sequential model estimation, assessing changes in direct effects when trust and confidence variables were introduced, rather than through causal path modeling or bootstrapped indirect effects. Reliability across constructs was strong (PSI α = .88; PER α = .86; TRU α = .90; DCF α = .87; INT α = .85). Mean scores exceeded the scale midpoint for all constructs (PSI M = 3.88; PER M = 3.61; TRU M = 3.74; DCF M = 3.69; INT M = 3.58). Correlation and regression results indicated that PSI and PER were positively associated with trust (PSI β = .41, p < .001; PER β = .34, p < .001; R^2 = .52). Trust accounted for substantial variance in decision confidence (β = .49, p < .001; DCF model R^2 = .57), and decision confidence accounted for substantial variance in intention to rely on AI recommendations (β = .43, p < .001; INT model R^2 = .49). When trust and confidence were included in the intention model, the direct associations of PSI and PER with intention were no longer statistically significant, indicating indirect relationships operating through trust and confidence. Overall, the findings suggest that SHAP interpretability and explanation robustness are associated with trust formation and confidence calibration, which together account for meaningful variance in reliance intentions in high-stakes healthcare and finance decision-support contexts.

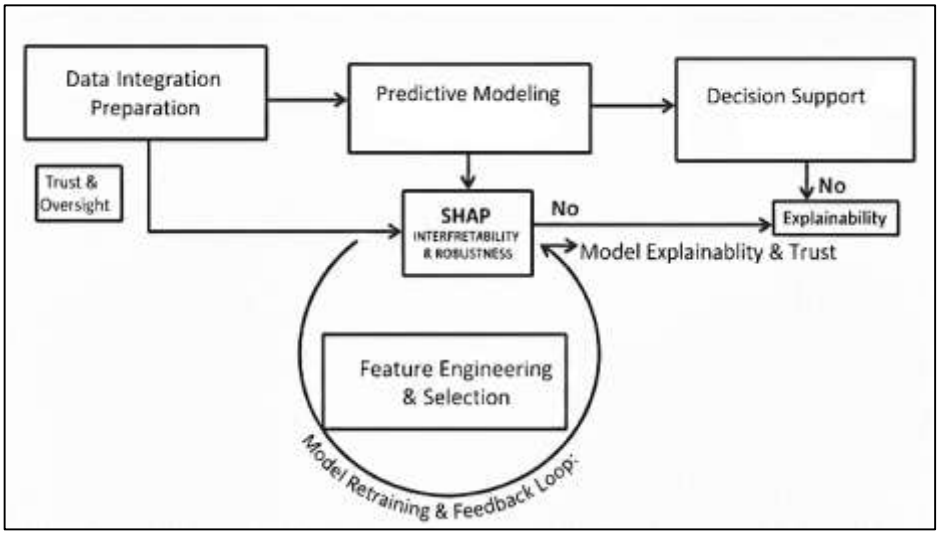
Keywords

SHAP Explainability, Interpretability, Explanation Robustness, Trust in AI, Decision Confidence.

INTRODUCTION

Artificial intelligence (AI) is commonly defined as the design and use of computational systems that perform tasks associated with human cognition, such as pattern recognition, prediction, classification, and decision recommendation in complex environments. Within AI, machine learning (ML) refers to algorithmic methods that learn statistical relationships from data to generate predictive models that generalize beyond observed samples, often outperforming rules-based approaches when the input space is high-dimensional and nonlinear (Dietvorst et al., 2015). In high-stakes decision support, ML models are used to provide risk scores, triage priorities, diagnostic probabilities, fraud likelihood estimates, or credit-default probabilities that shape real-world outcomes across healthcare and finance. The "high-stakes" label is used for contexts where errors carry substantial costs, including patient harm, inequitable access to care, regulatory breaches, financial losses, and systemic instability. The rapid diffusion of ML into high-stakes workflows has intensified attention to interpretability and explainability, because decision makers and overseers frequently require reasons, not only predictions. Surveys and syntheses characterize explainable AI (XAI) as a family of methods and governance practices that aim to make ML outputs intelligible to humans by representing how inputs contribute to predictions, how models behave across populations, and how uncertainty or instability enters the decision pipeline (Akhtar & Mian, 2018).

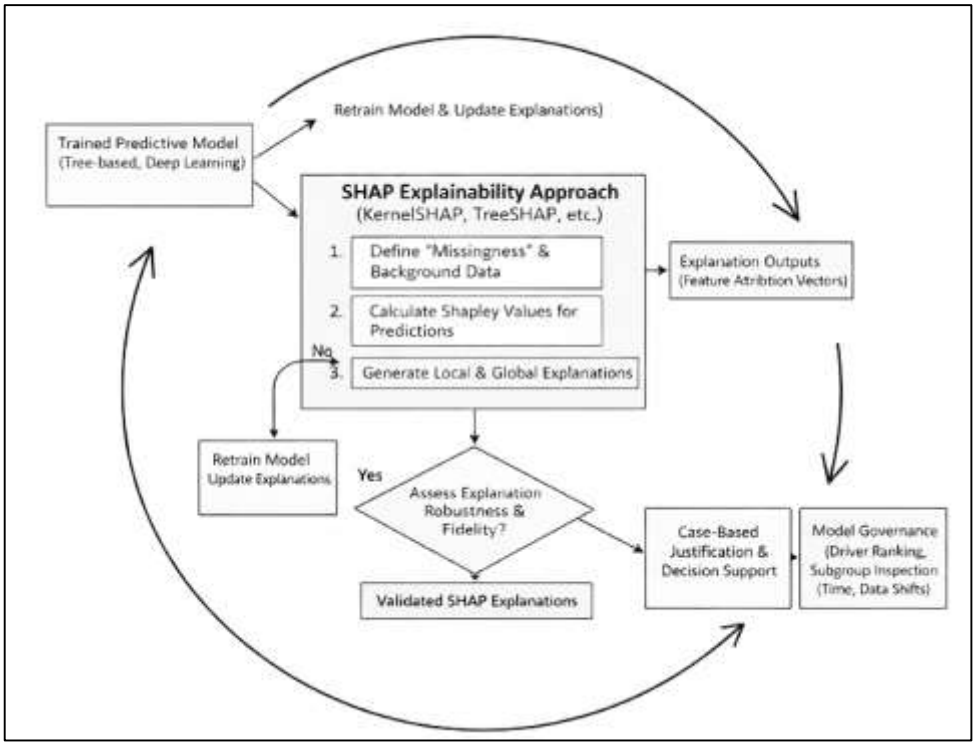
Figure 1: XAI Governance in High-Stakes Decision Support Systems



SHAP (SHapley Additive exPlanations) is a widely used explainable AI approach that operationalizes feature attribution using Shapley values from cooperative game theory, allocating each feature a contribution to an individual prediction under axioms such as efficiency, symmetry, and additivity. In practice, SHAP is valued because it provides local explanations (instance-level attribution vectors) that can be aggregated into global summaries of feature influence, supporting both case-based justification and model governance activities such as driver ranking, subgroup inspection, and monitoring of attribution drift over time (Hoff & Bashir, 2015). Methodologically, SHAP includes model-agnostic estimation procedures (e.g., KernelSHAP) and model-specific variants that exploit structure for efficiency (e.g., TreeSHAP), which has contributed to its adoption in enterprise workflows where explanation artifacts must be generated at scale and reported in standardized formats. However, the literature also emphasizes that SHAP explanations depend on how "missingness" and the background distribution are defined, meaning that attributions can change with different reference datasets, preprocessing choices, and sampling approximations, which raises reproducibility and robustness concerns in high-stakes settings (Miller, 2019). A central technical limitation concerns correlated features: when independence assumptions are used, SHAP can implicitly evaluate unrealistic coalitions that violate the data manifold, potentially producing misleading attributions that appear coherent but do not reflect plausible domain states; dependence-aware alternatives and conditional estimation

approaches have been proposed to address this issue. Related scholarship further argues that post hoc attributions should not be treated as guarantees of model validity, and that explanation outputs require evaluation for fidelity, stability, and susceptibility to manipulation, particularly when explanations are used for audit, compliance, or decision justification. Collectively, this literature positions SHAP as a practically influential attribution method whose governance value is strongest when explanation configuration is documented and explanation robustness is explicitly assessed alongside predictive performance.

Figure 2: SHAP-based Explainable AI Workflow



Morever, Conceptualizations of interpretability in applied ML also emphasize that "understanding" is not a single property: it depends on audience, task, stakes, and the degree to which an explanation enables humans to evaluate validity, fairness, and reliability within operational constraints (Murdoch et al., 2019). In internationally deployed systems, the pressure for explainability is amplified by cross-jurisdictional requirements for accountability, auditability, and documentation across heterogeneous institutions, languages, and standards of professional practice. In this environment, explainability becomes a core component of decision support quality, because it interacts with the credibility of evidence, the defensibility of actions, and the transparency of automated recommendations. The research community therefore treats XAI not as a cosmetic addition but as an integral element of safe and accountable use of ML in domains where decisions have enduring consequences for individuals and institutions.

Explainability is often discussed alongside interpretability, transparency, and human trust, yet these constructs operate at different levels of analysis. Interpretability is typically framed as the extent to which a human can understand the internal mechanics of a model or its mapping from inputs to outputs, while explainability is frequently operationalized as the provision of human-consumable reasons for specific predictions or global model behavior. Scholars in AI and the social sciences show that explanations are evaluated using human reasoning patterns, including causal narratives, contrastive "why this outcome" structures, and context-sensitive relevance judgments rather than purely mathematical completeness (Miller, 2019). Human–AI interaction research adds that explanations influence reliance and calibration: users must decide when to accept automated advice and when to override it. Empirical syntheses of trust in automation identify multi-layered drivers of reliance, including perceived competence, predictability, transparency, and the alignment of system behavior with user goals and norms (Hoff & Bashir, 2015). Behavioral evidence also indicates that

people sometimes avoid algorithmic advice after witnessing errors ("algorithm aversion"), while other settings show preference for algorithmic judgments ("algorithm appreciation"), highlighting that the same technology can produce divergent reliance patterns depending on task framing, accountability conditions, and perceived controllability (Dietvorst et al., 2015). Studies focused on explanation styles demonstrate that explanation format and level of detail shape perceived fairness, perceived usefulness, and trust, and that design choices can support more accurate trust calibration by helping users detect limitations rather than merely increasing confidence (Naiseh et al., 2023). Complementary work on trust calibration further indicates that the objective is not maximized trust but appropriately matched trust, because over-reliance and under-reliance each degrade decision quality in human–AI collaboration (Okamura & Yamada, 2020). In high-stakes contexts, these findings carry international significance because decision makers often operate within institutional protocols that demand both interpretability for audit and communicability for clients, patients, and regulators. Consequently, explainability is treated as a socio-technical property that links model behavior to human judgment, organizational accountability, and the traceability of decisions across diverse stakeholders.

Among XAI methods, Shapley-value attribution has become prominent because it offers a principled way to allocate credit for a prediction across input features under a cooperative game-theoretic framing. In operational ML, Shapley-based approaches are frequently implemented through SHAP (SHapley Additive exPlanations) variants that produce local explanations for individual predictions and can be aggregated into global summaries of feature influence. A widely cited application in tree-based models demonstrates how local explanation vectors can be combined to form global understanding of model structure and feature effects, supporting both case-level interpretability and population-level review in nonlinear predictive systems (Obermeyer et al., 2019). This local-to-global bridge is particularly relevant to high-stakes decision support because institutions often need both micro-level justification ("why this patient is high risk" or "why this applicant is rejected") and macro-level governance ("which variables drive outcomes overall," "how drivers vary across subgroups," and "whether patterns are stable"). XAI surveys also document that additive feature attribution is attractive because it can align with conventional statistical reasoning used in regulated domains, where coefficient-based interpretations are familiar and explanation artifacts can be incorporated into reports and oversight processes (Arrieta et al., 2020; Bussmann et al., 2021). At the same time, scholarship distinguishes explanations of black-box models from inherently interpretable models and argues that explanation layers can be misunderstood as guarantees of model validity, particularly under high stakes where users seek reassurance rather than critical evaluation (Rudin, 2019). Technical reviews of deep network interpretation further show that explanation methods vary in assumptions and failure modes, underscoring the need to test explanation fidelity, stability, and sensitivity under realistic data conditions rather than treating any explanation output as self-validating (Reddy, 2022). SHAP is therefore situated at the center of a broader methodological tension: it provides actionable interpretive artifacts at scale, yet its practical value depends on whether the explanation remains consistent under perturbations, sampling variation, correlated features, and domain shifts. This tension motivates robustness testing, including stability checks across resampling schemes, stress tests under noisy or adversarial inputs, and comparative evaluation across explanation methods for the same predictive task.

Healthcare provides a canonical setting for high-stakes ML because model outputs can influence diagnosis, prognosis, resource allocation, and treatment decisions across hospitals and health systems. Clinical ML research describes how modern models leverage electronic health records (EHR), imaging, and longitudinal data streams to generate predictions that support triage and risk stratification, while also noting that translation into routine care requires careful evaluation, monitoring, and integration into clinical workflows (Rajkomar et al., 2018). High-profile demonstrations of deep learning in medical imaging illustrate the promise of ML for pattern recognition at scale, but they also intensify scrutiny of transparency because clinicians must justify actions, communicate reasoning to patients, and document decisions in regulated environments (Esteva et al., 2017). In practice, clinicians and health organizations face a dual accountability problem: they must assess model accuracy and also defend model-driven recommendations using explanations that are clinically meaningful. Commentary in medical literature frames ML as a tool that reshapes how evidence is extracted from large datasets, with clinical benefit

depending on rigorous validation and appropriate use conditions (Gramegna & Giudici, 2021). At the same time, health equity research documents that widely used prediction algorithms can encode harmful biases, producing unequal access to supportive care even when the stated objective appears neutral, making transparency and auditing central requirements for responsible deployment (Guidotti et al., 2018). XAI-focused healthcare scholarship adds that explanation is not merely a technical artifact; it is a multidisciplinary challenge involving ethics, clinical reasoning, accountability, and the practical constraints of time-limited decision making (Amann et al., 2020). Work in medical AI also underscores the need to distinguish explanation from justification, because a plausible explanation can coexist with inadequate calibration, poor generalization, or unintended proxy effects. Related discussions in digital health emphasize that explainability is intertwined with governance, documentation, and stakeholder communication in healthcare ecosystems where decisions are audited by multiple parties and standards differ across regions (Amann et al., 2022). These dynamics make healthcare an essential domain for examining SHAP interpretability and robustness testing within decision support pipelines that must satisfy both predictive performance expectations and socio-technical accountability requirements across varied institutional contexts.

Although explainable AI has received substantial attention in both healthcare and finance, a key gap remains in how explanation *robustness* is evaluated and interpreted in practice. Existing studies frequently assess explanation methods in technical terms (e.g., fidelity, feature attribution behavior under perturbation, sensitivity to correlated inputs) or examine user-centered outcomes such as perceived transparency, trust, and acceptance. However, there is a scarcity of research that directly connects the *technical stability* of explanation outputs to users' *perceptions of that stability* in high-stakes decision settings. In regulated environments, stability is not only a methodological concern; it is also an experiential property that shapes whether explanation artifacts are interpreted as reliable, defensible, and appropriate for accountability-oriented decisions. When stability is not explicitly measured and linked to user perceptions, organizations may assume that providing explanations is sufficient even if the explanation outputs vary meaningfully across resampling, minor input perturbations, or routine model updates. This unresolved linkage motivates empirical designs that evaluate robustness as a measurable technical property while also measuring whether stakeholders perceive those explanations as stable and whether such perceptions are associated with trust and confidence outcomes under high-stakes conditions.

This study examines explainable machine learning as a decision-support mechanism in high-stakes environments by focusing on SHAP-based interpretability and the robustness of explanation outputs under controlled testing conditions in healthcare and finance. The first objective is to operationalize and measure stakeholders' perceptions of SHAP interpretability, emphasizing clarity, understandability, and perceived usefulness of explanation artifacts when applied to realistic decision scenarios where accountability and accuracy are salient. The second objective is to assess the robustness of SHAP explanations through systematic stability testing, treating explanation reliability as a measurable property that can be evaluated by observing how feature attributions vary across structured perturbations to inputs, sampling conditions, and model configurations. The third objective is to examine the associations of perceived interpretability and perceived robustness with trust in AI-driven decision support, decision confidence during case evaluations, and intention to rely on AI recommendations within each domain case context. The fourth objective is to estimate the strength and direction of these relationships using quantitative modeling aligned with cross-sectional survey data, while accounting for individual differences in expertise, professional experience, and prior AI exposure. The fifth objective is to compare the healthcare and finance contexts to determine whether interpretability and robustness are evaluated similarly across domains or whether domain-specific accountability pressures are associated with different patterns of trust formation and reliance-related outcomes. The sixth objective is to specify these expectations as empirically testable hypotheses and to evaluate them using descriptive statistics, reliability analysis, correlation analysis, and multiple regression models applied to Likert-scale constructs. The final objective is to integrate technical robustness assessment and human perception measurement within a unified framework that produces a coherent dataset linking explanation behavior and stakeholder evaluation, supported by clearly defined constructs and measurable indicators suitable for quantitative testing.

LITERATURE REVIEW

The literature on artificial intelligence–driven decision support in high-stakes domains emphasizes that predictive performance alone is insufficient when model outputs influence consequential clinical and financial outcomes, because decision makers must justify, audit, and communicate the rationale behind recommendations in environments shaped by regulation, professional accountability, and risk governance. Within this body of work, explainable AI (XAI) is positioned as a set of methods and design principles intended to increase transparency, interpretability, and human understanding of model behavior, thereby supporting appropriate reliance and defensible decision making. Research across human–AI interaction and responsible AI governance further shows that explanation is not merely a technical add-on but a socio-technical mechanism that interacts with trust, perceived fairness, perceived usefulness, and perceived controllability, each of which affects acceptance and reliance on automated advice. In parallel, methodological scholarship distinguishes between intrinsically interpretable models and post hoc explanation techniques applied to complex models, warning that post hoc explanations can be misunderstood as guarantees of correctness if their limitations are not evaluated and communicated. Among post hoc approaches, Shapley-value feature attribution has gained prominence through SHAP-based methods that provide local explanations for individual predictions and can be aggregated into global summaries of feature influence, offering a practical pathway for explanation delivery in real decision-support workflows. However, an increasingly important theme in recent literature is that explanation quality must be assessed beyond informativeness and visual appeal, because explanation outputs may be sensitive to correlated variables, sampling variation, preprocessing choices, and model retraining—conditions that can produce instability in feature attributions even when predictive accuracy remains stable. This has motivated a growing emphasis on robustness testing of explanations, where stability and consistency are treated as measurable properties using perturbation-based methods, resampling schemes, or retraining variations to examine whether explanations remain dependable under realistic uncertainty. Healthcare-focused studies highlight additional pressures related to clinical accountability, patient communication, and bias risk, while finance-focused studies emphasize auditability, compliance requirements, rare-event modeling challenges, and the operational consequences of distribution shifts. Taken together, the literature suggests that a rigorous evaluation of explainable machine learning for high-stakes decision support requires integrating technical assessment of SHAP explanation robustness with user-centered measurement of interpretability perceptions and trust-related outcomes, enabling empirical examination of how explanation stability and perceived clarity jointly shape confidence and reliance across domain case contexts.

AI-Driven Decision Support in High-Stakes Domains

AI-driven decision support in high-stakes domains refers to the use of computational prediction and classification systems to inform decisions where errors can trigger serious harm, legal exposure, or systemic loss. In healthcare, these systems ingest clinical histories, laboratory values, imaging outputs, and workflow signals to generate risk estimates that shape triage, diagnosis support, and resource allocation across diverse care settings. Decision support is embedded in time-constrained, multidisciplinary workflows, so tools must deliver outputs that are timely, comprehensible, and compatible with documentation requirements. The literature describes how data scale and model complexity have enabled predictive performance gains, while also intensifying the need for careful evaluation of clinical validity, subgroup performance, and operational fit because clinicians remain accountable for actions taken on the basis of model outputs (Beam & Kohane, 2018). Translation from prototype to routine use also depends on how risk scores and explanations are presented to frontline users, since poorly communicated model information can produce misunderstanding, miscalibration, and inconsistent uptake across roles and care settings. Operational deployment adds additional constraints, including interoperability with electronic records, governance over model updates, and mechanisms for tracking performance and user feedback after release. To address these realities, research emphasizes structured approaches to communicating model purpose, inputs, limitations, and intended use conditions in formats that clinicians can quickly interpret within clinical context (Sendak et al., 2020). Across international health systems, these themes converge on a shared requirement: AI decision support must be evaluated not only for accuracy but also for its capacity to support

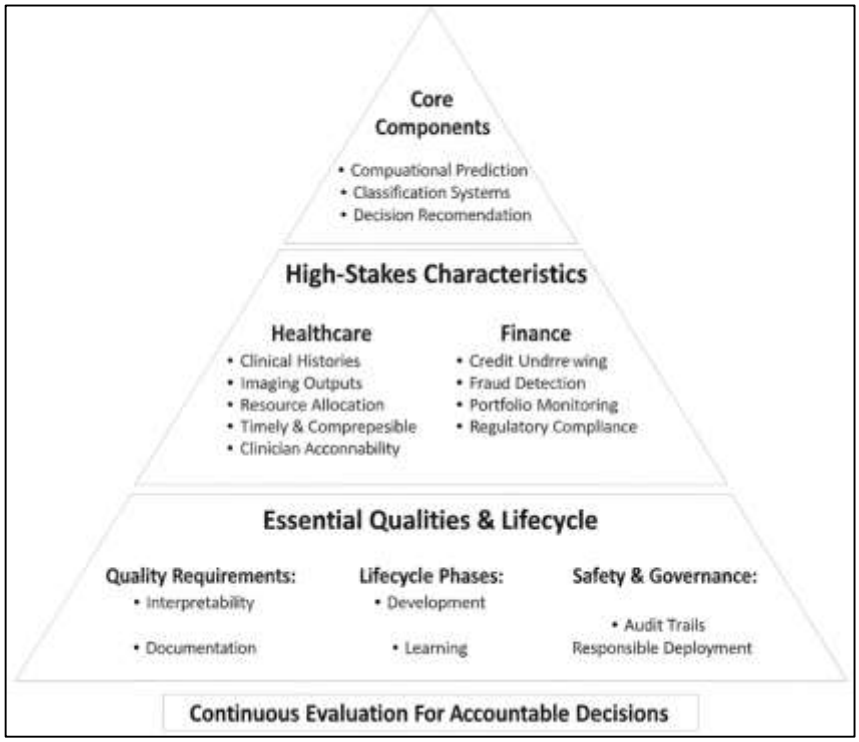
accountable decisions under real workflow pressures, variable data quality, and heterogeneous patient populations (Jinnat & Kamrul, 2021; Ashraful et al., 2020). Consequently, high-stakes healthcare decision support is increasingly framed as a lifecycle problem that spans development, validation, implementation, monitoring, and organizational learning, with interpretability and robustness becoming essential qualities for sustained use. This framing underpins domain-wide adoption and oversight (Fokhrul et al., 2021; Towhidul et al., 2022).

High-stakes decision support raises questions about safety engineering and accountability mechanisms that surround AI-enabled tools when recommendations can influence treatment pathways, escalation decisions, or access to scarce resources (Faysal & Bhuya, 2023; Hammad & Mohiul, 2023). Clinical decision support has historically relied on transparent logic, yet machine-learned systems can change through retraining, data refreshes, and interface updates, creating the possibility that similar inputs yield different outputs over time. The literature therefore distinguishes one-time validation from ongoing assurance, emphasizing processes for verification, certification, monitoring, and incident reporting that operate alongside clinical quality and safety programs (Masud & Hammad, 2024; Md & Praveen, 2024). Consensus-oriented work synthesizes practical building blocks for responsible AI-enabled clinical decision support, including documentation of model intent, explicit statements of autonomy and human oversight, continuous performance surveillance, and standardized pathways to report unexpected behavior or harm (Labkoff et al., 2024). This governance perspective connects to interpretability because explanation artifacts are used in training, review, and audit conversations, and to robustness because stability under operational change affects both trust and defensibility. Implementation research also notes that AI decision support is consumed by multiple stakeholders clinicians, informatics teams, compliance officers, and patients each needing different explanation granularity to judge whether a recommendation fits a case. Responsible deployment depends on aligning technical evaluation with human-centered evaluation of usability and comprehension, and on ensuring that model updates do not silently shift decision logic without review (Newaz & Jahidul, 2024; Sai Praveen, 2024). Internationally, this alignment is complicated by varied legal regimes, data standards, and clinical practice norms, so governance frameworks emphasize transparency of data provenance, clarity about intended populations, and plans for managing drift when patient mix or care processes change. High-stakes clinical decision support is thus treated as an institutional capability that must be maintained, audited, and learned from continuously (Faysal & Aditya, 2025; Azam & Amin, 2024).

In finance, AI-driven decision support is applied to credit underwriting, portfolio monitoring, fraud detection, and early warning analytics, where model outputs influence access to capital, pricing, loss mitigation actions, and supervisory responses. Because these decisions scale across large populations, small shifts in predictive performance can translate into material changes in aggregate loss, consumer outcomes, and institutional risk exposure (Hammad & Hossain, 2025; Towhidul & Rebeka, 2025). Empirical research shows that machine-learning methods can improve consumer credit-risk forecasting by combining credit bureau attributes with transaction signals to produce timely, nonlinear risk scores, and it links these forecasts to operational interventions such as credit-line adjustments and portfolio-level risk monitoring (Khandani et al., 2010; Yousuf et al., 2025; Azam, 2025). The high-stakes character of these applications arises from the economic consequences of misclassification and the legal and reputational costs associated with decisions that are difficult to justify. FinTech lending research also shows that technology-enabled interfaces can compress decision timelines and broaden data sources used in screening, which increases the need for governance structures that can explain outcomes to customers and regulators (Berg et al., 2022; Tasnim, 2025; Zaheda, 2025b). Across jurisdictions, financial institutions operate under model-risk management expectations that require documentation, independent review, and change control for models that inform material decisions; explainability methods are used to summarize drivers of risk scores and generate reason codes for adverse actions (Zaheda, 2025). Robustness is central because macroeconomic conditions, borrower behavior, and fraud strategies shift over time, so models can deteriorate in production and produce unstable attributions that complicate auditing. These pressures position finance as a complementary domain to healthcare for studying explainable machine learning, enabling comparison of explanation needs under different data-generating processes and accountability routines. Financial decision

support frequently interacts with automated workflows, so explanation artifacts must travel across internal teams, customer communications, and supervisory review. This multi-audience requirement elevates interpretability and stability as measurable qualities rather than optional reporting features in practice.

Figure 3: Triangle Framework For AI-Driven Decision Support In High-Stakes Domains

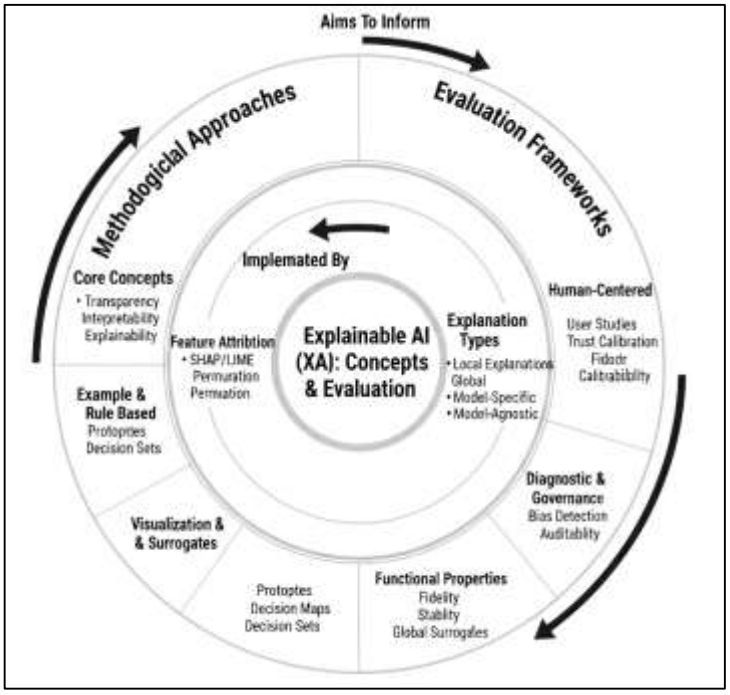


Explainable AI (XAI) Concepts, Approaches, and Evaluation

Explainable artificial intelligence (XAI) addresses the reality that many high-performing machine-learning models operate as complex function approximators whose internal representations are not readily interpretable by human decision makers. In high-stakes settings, explanation functions as an interface between statistical inference and accountable action, connecting model behavior to the information needs of clinicians, analysts, auditors, managers, and affected individuals. Core concepts distinguish transparency, interpretability, and explainability. Transparency describes the degree to which system components, data provenance, and processing steps are open to inspection. Interpretability refers to the extent a human can understand a model's mapping from inputs to outputs, either because the model is simple by design or because outputs can be decomposed into comprehensible parts (Selvaraju et al., 2020). Explainability emphasizes the communication of reasons for predictions or decisions in forms that support questioning and verification within a task context. These definitions guide research design by treating explanations as purposeful artifacts evaluated relative to audience goals and domain conventions rather than as generic visualizations. Because stakeholders differ in what they must justify, XAI research separates local explanations, which justify a single prediction, from global explanations, which describe behavior across cases. A complementary distinction separates model-specific explanations, which exploit structure such as gradients or tree paths, from model-agnostic explanations, which use query access to approximate behavior. This taxonomy supports comparison of methods and encourages evaluation designs that connect explanation form to decision tasks, governance, and user comprehension, while emphasizing the documentation of assumptions because explanation outputs depend on data processing, feature definitions, and the pipeline that consumes them (Mohseni et al., 2020). Internationally, these distinctions matter because institutions must align technical explanations with legal duties, professional standards, and expectations for reason-giving, which vary across jurisdictions and organizational cultures.

In addition to interpretability, high-stakes evaluation requires an explicit account of explanation robustness, because explanation outputs can vary across repeated runs and operational changes even when predictive performance remains stable. In this study, Explanation Robustness is defined as the degree to which an explanation remains stable and reproducible under small, plausible perturbations to (a) input values, (b) sampling/background reference sets, and (c) model or pipeline conditions (e.g., retraining, resampling, or configuration choices). Operationally, robustness concerns whether the identity, ranking, and directional influence of the most influential features in an explanation remain consistent when explanation generation is repeated under controlled perturbation and resampling procedures. Under this definition, explanation robustness is treated as a measurable quality attribute of explanation artifacts rather than as a general belief about AI reliability. This distinction is important in high-stakes contexts because unstable explanations can weaken auditability, defensibility, and users' capacity to justify decisions to stakeholders, regulators, or oversight bodies.

Figure 4: Conceptual Framework For Explainable AI (XAI) Concepts



SHAP Interpretability in Practice

SHAP interpretability is commonly operationalized in practice as a feature-attribution workflow that converts a model prediction into an additive set of contributions assigned to input variables, enabling users to inspect which inputs increased or decreased a specific risk score and to summarize driver patterns across many cases. In applied deployments, SHAP outputs are used in two complementary ways: local explanations support case-level review by listing feature contributions for a single patient record or financial application, while aggregated explanations summarize overall model behavior by ranking features, profiling attribution distributions, and comparing driver patterns across cohorts or time windows. These practices align with governance needs because they create artifacts that can be stored, reviewed, and communicated as part of model documentation and quality assurance routines. In real settings, however, the practical meaning of a SHAP attribution depends on how "missingness" or feature removal is defined when computing marginal contributions, because the method must represent what the model would output if a feature were unknown. When features are correlated, naive independence assumptions can generate unrealistic synthetic samples during the explanation procedure, which can distort attributions and produce narratives that do not align with plausible domain conditions. A major practical implication is that SHAP explanations may appear coherent while encoding counterfactual input combinations that would not occur in real patients or real borrowers, undermining explanation credibility for expert users. Research extending Kernel SHAP to dependent

features formalizes this problem and proposes alternatives that better respect feature dependence, showing that dependence-aware estimation changes attribution patterns and reduces misleading explanations under correlation structures that are typical in applied data (Aas et al., 2021). In practice, this dependence issue is not an edge case, because healthcare measurements often co-move through physiology and treatment protocols, and finance attributes frequently co-vary through income, utilization, and portfolio structure. As a result, practitioners increasingly treat SHAP interpretability not as a single output but as a process that requires careful alignment between data-generating realities and the mathematical assumptions embedded in the explanation algorithm. When SHAP explanations vary across repeated runs, small input changes, or routine model updates, users may notice changes in which features are highlighted, the rank order of "top drivers," or the direction of feature contributions, even when the predicted score remains similar. In high-stakes workflows, these visible shifts can be interpreted as inconsistency in the system's rationale, reducing perceived defensibility for audit, documentation, or stakeholder communication. As a result, explanation stability functions as a reliability cue at the interface level: explanations that appear consistent are more likely to be perceived as dependable, whereas explanations that "change" can be associated with lower trust and reduced decision confidence, particularly when accountability for the final decision remains with the human user.

Figure 5: Triangle Framework For SHAP Interpretability In Decision Support

A second practical dimension of SHAP use concerns computational constraints and repeatability when explanations are produced at scale or under strict latency requirements. High-stakes decision support often involves large populations, frequent scoring, and the need to regenerate explanations after model updates, data refreshes, or monitoring alerts. These operational demands raise questions about which forms of SHAP can be computed efficiently and which settings are inherently intractable. Complexity analysis shows that computing SHAP explanations can be computationally difficult even for modeling settings that are widely used in practice, and it clarifies that tractability depends on both the model class and the assumptions about the input distribution used in the explanation procedure (Van den Broeck et al., 2022). From an implementation standpoint, this matters because explanation pipelines must be stable, auditable, and reproducible across runs, and because governance teams need consistent artifacts to compare pre-deployment and post-deployment behavior. If explanation generation is

computationally expensive, institutions may resort to approximations, sampling, or caching strategies, which can introduce additional variability into the explanations presented to end users. Variability becomes a practical risk when stakeholders interpret explanation changes as evidence that the model is "changing its mind," even if predictive performance remains similar. Therefore, practitioners often implement SHAP under explicit operational policies, such as fixed random seeds, controlled background datasets, and standardized preprocessing steps, so that attribution outputs are comparable across time. In healthcare, repeatability supports clinical review meetings and audit trails for adverse events; in finance, repeatability supports model risk management functions that compare explanation distributions across segments and reporting periods. These constraints also motivate robustness testing protocols that measure whether explanation rankings and directions remain consistent under controlled variation, treating explanation stability as a measurable operational property rather than an informal expectation.

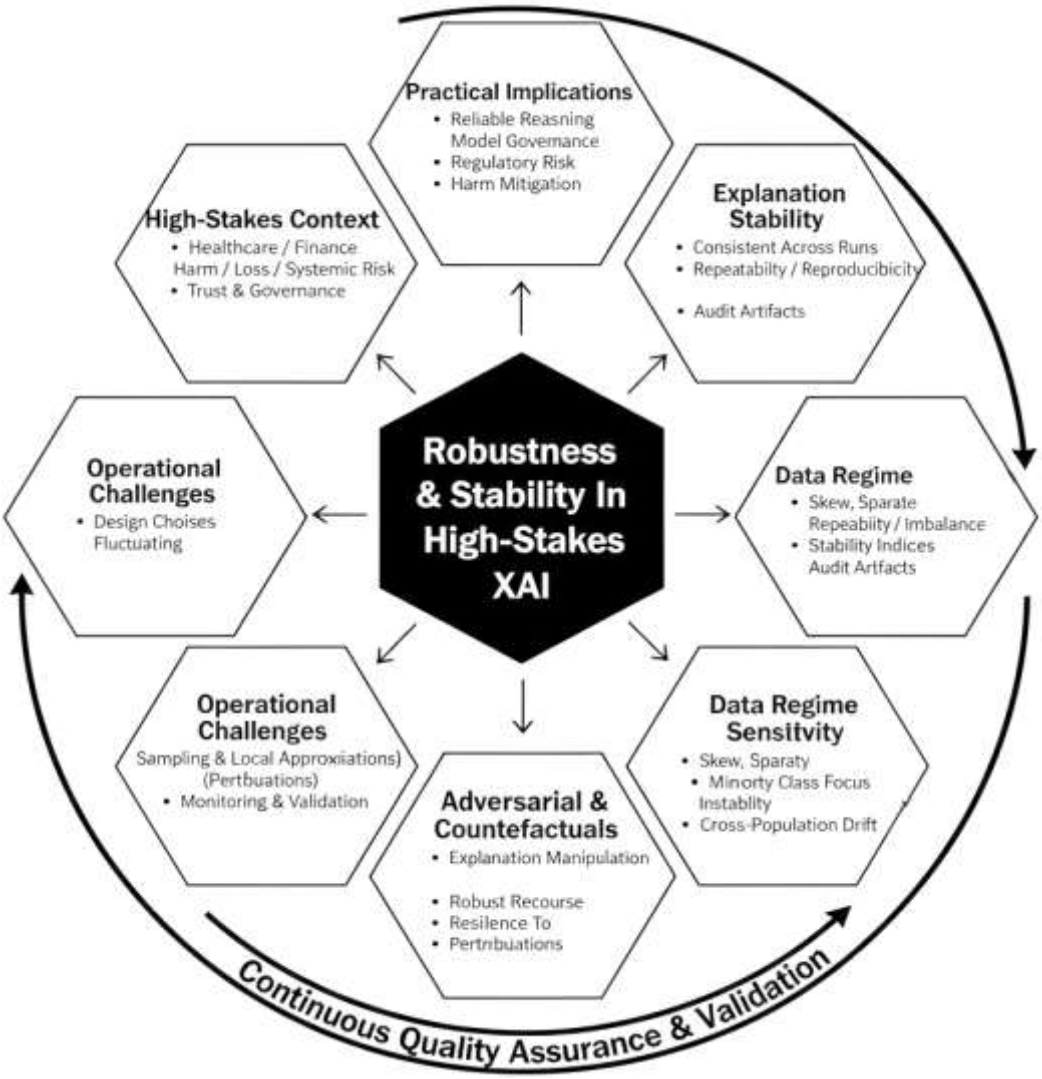
A third practical theme is that SHAP interpretability is increasingly used alongside sensitivity-analysis perspectives that distinguish local, global, and hybrid "glocal" understandings of model influence. In applied decision support, local attributions help justify an individual decision, yet governance often requires global understanding of which variables drive outcomes overall and whether interactions or dependence structures alter that story. Shapley effects in global sensitivity analysis provide a theoretical bridge by interpreting feature importance through variance contributions, supporting explanations that remain well-defined even when inputs are dependent and interactions are present (Song et al., 2016). Recent work emphasizes that multiple Shapley-based formulations can yield complementary insights at different scales, motivating practice-oriented frameworks where local case explanations are interpreted alongside global summaries and interaction indices rather than treated as substitutes (Borgonovo et al., 2024). At the same time, practice-facing research highlights that conventional SHAP implementations may misinterpret correlated, high-dimensional inputs because they can produce coalitions that violate the data manifold, leading to attributions that do not reflect realistic feature configurations. Manifold-based Shapley approaches propose generating Shapley values in a latent representation that preserves correlations and then mapping attributions back to the original feature space, aiming to correct misinterpretations and improve feasibility in complex high-dimensional settings (Hu et al., 2024). For high-stakes domains, these developments matter because they connect explanation validity to the structure of the data, reinforcing that explanation robustness requires both algorithmic stability and distributional realism. Practitioners therefore increasingly combine SHAP outputs with dependence diagnostics, subgroup profiling, and robustness checks to ensure that explanations used for justification, audit, and communication remain consistent with the operational realities of healthcare and finance decision pipelines.

Robustness and Stability of Explanations in High-Stakes XAI

Robustness in explainable machine learning refers to the extent to which an explanation remains consistent, meaningful, and decision-relevant when an input record (or the explanation procedure itself) is exposed to small, plausible perturbations. In high-stakes decision support, explanations are not simply interpretive accessories; they function as audit artifacts that justify decisions, communicate risk, and document accountability. As a result, explanation instability can damage trust and governance even when a model's predictive accuracy appears stable. A core reason is that many post-hoc explanation methods rely on sampling, local approximation, and design choices (e.g., perturbation distributions, neighborhood size, background data) that can yield materially different feature-importance rankings across repeated runs. Robustness therefore needs to be framed as a repeatability and reproducibility requirement: if the same case is explained multiple times under equivalent conditions, the explanation should converge to a similar set of influential features and comparable contribution magnitudes. A practical way to operationalize this requirement is through explicit stability indices that quantify the variability of selected features and their associated weights across repeated explanations. For example, stability-focused work on LIME in risk-oriented settings formalizes complementary indices to capture both variation in the identity of explanatory features and variation in the coefficients assigned to those features, enabling practitioners to detect when explanations are effectively "moving targets" rather than reliable decision rationales (Visani et al., 2022). Such indices are particularly valuable in healthcare and finance because decisions are often

reviewed after the fact by oversight bodies, and inconsistent explanations can be interpreted as evidence of unreliable reasoning, weak model governance, or inadequate validation. Robustness testing, accordingly, becomes a necessary component of explanation quality assurance, alongside more familiar checks of predictive performance.

Figure 6: Hexagonal Framework For Robustness And Stability In High-Stakes XAI



A second robustness challenge is that explanation stability is shaped by the data regime in which models are trained and evaluated, meaning that explanation variance can increase when datasets exhibit skew, sparsity, or class imbalance. This issue is especially salient in finance (e.g., default prediction) and healthcare (e.g., rare adverse outcomes), where the minority class is frequently the decision-critical group that demands the clearest justification. When minority cases are scarce, the learned decision boundary may become sensitive to sampling fluctuations and local neighborhood composition, which can amplify instability in local explanations. In practical terms, the same individual case may receive different feature-importance rankings depending on whether the explainer's neighborhood is dominated by majority-class instances or contains sufficient minority-class representation to characterize the relevant decision surface. Empirical evidence from imbalanced credit-scoring contexts shows that the stability of both LIME- and SHAP-based interpretations degrades as class imbalance increases, indicating that interpretability itself can be adversely affected by the statistical structure of the dataset (Y. Chen et al., 2024). This has direct implications for high-stakes deployment because stability failures are likely to concentrate precisely where models are most scrutinized—borderline cases, minority outcomes, and high-loss decisions. Robustness therefore

should be assessed under multiple data conditions, including controlled imbalance levels, alternative resampling strategies, and realistic distribution shifts. In healthcare, analogous concerns arise when models trained in one hospital system or population subgroup are applied elsewhere: even if performance transfers moderately well, explanation stability may deteriorate because feature relationships and prevalence patterns differ, producing fluctuating attributions that confuse clinicians and complicate documentation. Robustness testing must therefore treat the dataset as part of the explanation system, not merely a backdrop for model training.

Robustness must also be considered under adversarial and counterfactual perspectives, because explanations can be manipulated or can fail to provide dependable recourse under uncertainty. In adversarial settings, the threat is not only that model predictions can be attacked, but also that explanation outputs can be altered to appear innocuous, compliant, or fair while the underlying decision logic remains problematic. A comprehensive survey of adversarial attacks and defenses in explainable AI synthesizes how data poisoning, model manipulation, and backdoor strategies can preserve nominal predictive behavior while changing explanatory behavior, undermining auditability and high-stakes governance (Baniecki & Biecek, 2024). In parallel, robustness concerns appear in the relationship between interpretability and adversarial examples: attackers can exploit instabilities in saliency or attribution patterns to either evade detection or cause misleading interpretation signals. Work leveraging saliency characteristics for adversarial example detection highlights that interpretability outputs can change systematically under attack, and that these changes can be used diagnostically—underscoring that explanation behavior is itself a security-relevant surface that must be tested (Wang & Gong, 2021). Finally, robustness is central to counterfactual explanations used for recourse: if a recommended change is fragile, small adverse perturbations beyond the user's control can invalidate the recourse pathway or make it far more costly than anticipated. Formal treatment of robust counterfactuals demonstrates that counterfactual recommendations are often not robust, and that incorporating robustness into the search process can yield recourse options that remain feasible under adverse perturbations (Virgolin & Fracaros, 2023). For high-stakes healthcare and finance, these strands converge into a unified requirement: robust explainability must address run-to-run stability, data-regime sensitivity, adversarial resilience, and recourse reliability, because explanation failures can translate into regulatory risk, operational risk, and harm to affected individuals.

Theoretical Framework

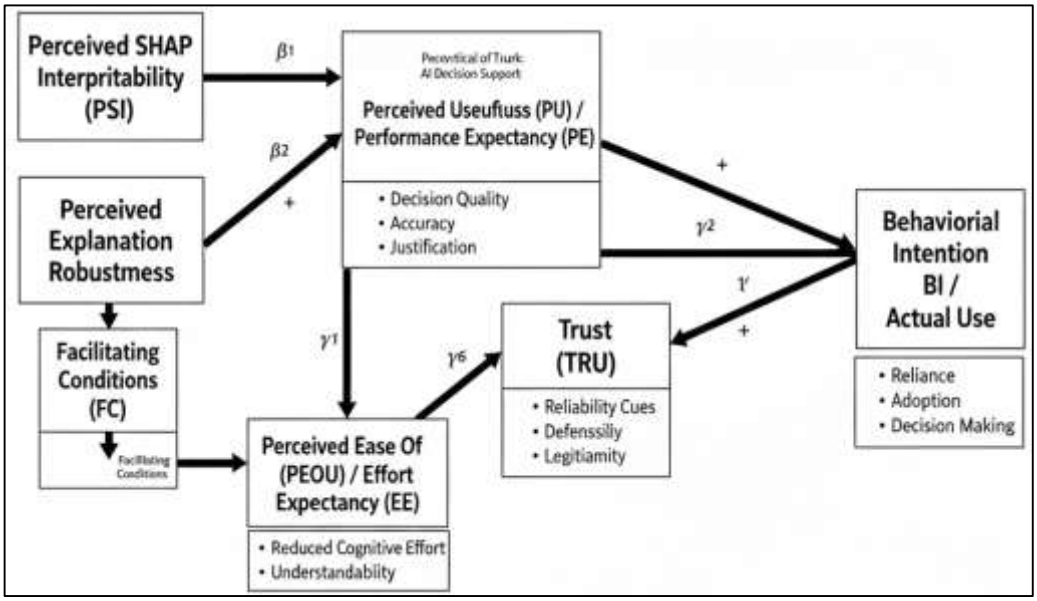
High-stakes AI decision support can be theorized as a special case of technology acceptance in which users evaluate not only usefulness and ease of use, but also the defensibility and reliability of acting on algorithmic recommendations under accountability constraints. In this study, **Technology Acceptance** Model 3 (TAM3) is used as the primary theoretical foundation to connect explanation quality perceptions to reliance-related outcomes, because TAM3 explains how belief formation regarding perceived usefulness (PU) and perceived ease of use (PEOU) accounts for variance in behavioral intention (BI) and use in organizational settings (Venkatesh et al., 2012). In high-stakes environments, PU is interpreted as decision performance expectancy, reflecting whether the system is perceived to improve decision accuracy, speed, and justification adequacy, while PEOU reflects the perceived cognitive effort required to interpret and apply model outputs. Within this TAM-consistent framing, perceived SHAP interpretability (PSI) is positioned as an information quality belief that is associated with stronger PU and lower interpretive burden because SHAP artifacts (e.g., feature attribution vectors, ranked drivers) can support intelligibility, verification against domain logic, and communication of rationale for audit and oversight purposes (Rai, 2020). The framework further treats Explanation Robustness as a distinct construct that must be explicitly defined rather than implied: in this study, explanation robustness refers to the degree to which explanation outputs remain stable and reproducible under small, plausible perturbations to input values, sampling or background reference data, and model or pipeline configurations (e.g., resampling, retraining, or parameter changes), with stability interpreted as consistency in the identity, ranking, and directional influence of influential features across repeated explanation generation. This robustness definition is intentionally scoped to explanation artifacts (not merely predictive stability), because in accountability-oriented domains, instability in explanation outputs can weaken defensibility even when predictive accuracy remains unchanged. To incorporate reliance mechanisms central to high-stakes decision support, the study

integrates trust as a mediating construct within a TAM-style acceptance structure, specifying that PSI and perceived explanation robustness (PER) are expected to be positively associated with trust in AI decision support (TRU), and that TRU is expected to account for variance in downstream decision confidence (DCF) and intention to rely/use (INT). Trust-in-automation theory is treated as complementary rather than competing with TAM3, because it clarifies why interpretability and robustness operate as different trust cues: interpretability primarily provides transparency cues that support mental-model formation, while robustness provides reliability cues that reduce concern that explanations are unstable or sensitive to minor changes in data or system state. Accordingly, the theoretical framing specifies regression-ready relationships consistent with the study's analytic approach, where TRU is modeled as a function of PSI and PER, and reliance-related outcomes are modeled as functions of trust and confidence, yielding a unified explanation of how explanation clarity and explanation stability perceptions are associated with trust calibration and reliance intentions in high-stakes healthcare and finance decision-support contexts.

Trust theory in the AI context clarifies why interpretability and robustness should be modeled as distinct antecedents. Trust research synthesizes evidence that AI differs from earlier deterministic technologies because it can adapt, learn, and behave in ways that are difficult for users to predict, making reliability signals and transparency cues central to trust calibration (Glikson & Woolley, 2020). In explainable ML decision support, SHAP interpretability primarily supplies *transparency cues* that help users form a mental model of why a recommendation was produced, while robustness testing supplies *reliability cues* that reduce concern that the explanation is unstable or sensitive to minor changes in data or model state.

Theoretical framing from information systems emphasizes that explainability objectives vary by stakeholder, and that explanations can serve governance, user decision making, and system improvement simultaneously, which supports modeling interpretability as a belief construct and robustness as an assurance construct rather than treating both as a single "transparency" variable (Meske et al., 2020). This separation is especially relevant in healthcare and finance where accountability processes require stable rationales across time and consistent decision justification across similar cases. As a result, the theoretical framework for this study treats SHAP interpretability (PSI) as a mechanism that supports understandability and perceived decision benefit, and explanation robustness (PER) as a mechanism that supports reliability and defensibility of explanations, with both influencing trust and downstream reliance-related outcomes in a measurable, testable structure aligned with regression-based hypothesis testing.

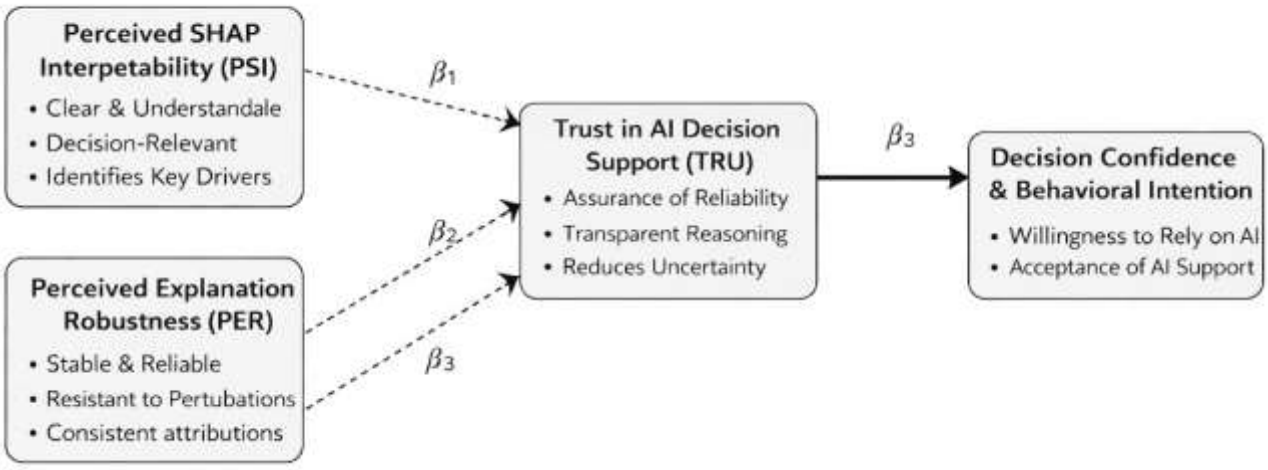
Figure 7: Theoretical Framework of the Study



Conceptual Framework

A conceptual framework for AI-driven explainable machine learning in high-stakes decision support must treat *explainability* as a measurable system property that influences human judgment through cognitive and behavioral mechanisms. In this study, Perceived SHAP Interpretability (PSI) represents the degree to which users judge SHAP-based explanations as clear, understandable, and decision-relevant, while Perceived Explanation Robustness (PER) represents the degree to which users believe those explanations remain stable under small, plausible perturbations. These beliefs are theorized to shape Trust Calibration (TRU) and downstream Decision Reliance/Decision Quality (DRQ) within healthcare and finance case contexts. The conceptual framing is grounded in the empirical observation that transparency can raise trust only when information is balanced and cognitively tractable, because excessive disclosure may overload users and reduce confidence in the system's outputs (Kizilcec, 2016). In this framework, PSI is not treated as "more information is always better"; rather, it is modeled as a quality perception that depends on whether explanations support intelligibility. Intelligibility research demonstrates that explanation form (e.g., "why" and "why-not") changes whether users can reason about system behavior and whether they can detect inconsistencies, which justifies PSI as a structured construct captured through multi-item Likert indicators rather than a single satisfaction item (Lim et al., 2009). Consequently, PSI is operationalized as the user's perceived ability to (a) identify key drivers, (b) connect drivers to domain logic, and (c) communicate reasons to stakeholders. PER is operationalized as the user's perceived confidence that the same case would yield a consistent attribution profile across repeated runs, background sets, and minor input changes. Together, PSI and PER function as the primary explanatory perceptions that feed into trust calibration and reliance outcomes within the proposed research model.

Figure 8: Conceptual Framework for AI-Driven Explainable Machine Learning



At the technical layer, the conceptual framework anchors interpretability to SHAP's additive attribution logic, which supports a transparent link between prediction and feature contributions. For a model $f(\cdot)$and an instance x, SHAP expresses the prediction as:

$$f(x) = \mathbb{E}[f(X)] + \sum_{i=1}^{M} \phi_i$$

where $\mathbb{E}[f(X)]$is the baseline expectation and $\phi_i$is the contribution of feature $i$across Mfeatures. Conceptually, PSI increases when users perceive the set $\{\phi_i\}$as coherent, sparse enough to interpret, and aligned with decision narratives. Robustness is represented by the stability of attribution vectors under repeated explanation conditions. A simple stability expression relevant for robustness testing is the correlation of attribution vectors between two runs:

$$S = \text{corr}\left(\phi^{(1)}, \phi^{(2)}\right)$$

where higher *S*indicates more consistent explanations for the same case. However, the human-centered evidence shows that interpretability cues do not automatically yield better teaming or better accuracy; explanations can increase acceptance of AI recommendations even when the AI is wrong, which makes *trust calibration* a critical mediating construct in high-stakes settings (Bansal et al., 2021). Complementing this, controlled experiments demonstrate that "more interpretable" models can produce counterintuitive user behaviors, including reduced error detection due to information overload, indicating that PSI must be measured distinctly from performance metrics (Poursabzi-Sangdeh et al., 2021). Therefore, the framework explicitly separates (a) explanation perceptions (PSI, PER), (b) psychological reliance mechanisms (TRU), and (c) outcome constructs (DRQ, overreliance risk). This separation supports quantitative testing using correlation and regression while preserving theoretical clarity between what explanations *are*, how they are *perceived*, and how they *affect* decisions. At the behavioral layer, the conceptual framework positions TRU as a mediator linking explanation perceptions to reliance-related outcomes in healthcare and finance case settings. The key concern is not only whether users *like* explanations, but whether explanations support calibrated reliance and improved decision handling under uncertainty. Evaluation research cautions that proxy tasks and subjective trust ratings can be misleading predictors of real task performance, implying that DRQ should be measured using decision-task aligned items (or scenario-based items) in addition to perception measures (Buçinca et al., 2020). Accordingly, the framework specifies three regression-ready relationships suitable for the study design: (1) $TRU = \alpha_0 + \alpha_1 PSI + \alpha_2 PER + \varepsilon$; (2) $DRQ = \beta_0 + \beta_1 PSI + \beta_2 PER + \beta_3 TRU + \varepsilon$; and (3) an optional overreliance risk model $OR = \gamma_0 + \gamma_1 TRU + \gamma_2 PSI + \gamma_3 PER + \varepsilon$, where OR captures tendencies to accept AI output without verification in high-stakes scenarios. These relationships map directly to the planned quantitative approach (descriptive statistics for constructs, correlation matrix, and multiple regression with hypothesis tests). PSI and PER are measured via Likert-scale constructs reflecting interpretability clarity and explanation stability perceptions; TRU is measured as calibrated confidence in system recommendations; and DRQ is measured as perceived decision quality, justification adequacy, and perceived error-detection capability within case vignettes. This conceptual framework therefore integrates SHAP's formal explanation structure with robustness notions and empirically grounded human response patterns, enabling hypothesis-driven testing in cross-sectional case-study contexts.

METHODS

The methodology for this study has been designed to examine explainable machine learning for high-stakes decision support by integrating a quantitative, cross-sectional survey with two domain-based case study contexts in healthcare and finance. The research approach has been structured to capture both the technical characteristics of SHAP explanations and the human-centered perceptions that have shaped trust and reliance on AI recommendations in consequential decision scenarios. A structured survey instrument was developed using a five-point Likert scale to measure the study constructs: perceived SHAP interpretability, perceived explanation robustness, trust in AI decision support, decision confidence, and intention to rely on AI outputs. Survey items were adapted and contextualized from established measurement scales in the information systems, human–AI interaction, and trust-in-automation literature, rather than created de novo. Specifically, interpretability- and explanation-related items were adapted from prior XAI and intelligibility research that operationalizes perceived understanding, clarity, and usefulness of explanations (e.g., Lim et al., 2009; Kizilcec, 2016; Mohseni et al., 2020), while trust-related items were adapted from validated trust-in-automation and trust-in-AI scales emphasizing reliability, predictability, and confidence in system outputs (Hoff & Bashir, 2015; Glikson & Woolley, 2020). Decision confidence and intention-to-rely items were adapted from technology acceptance and decision-support adoption studies grounded in TAM and related extensions, which operationalize confidence, reliance intention, and willingness to act on system recommendations in organizational settings (Venkatesh et al., 2012; Rai, 2020). All items were reworded to reflect the specific context of SHAP-based explanation artifacts and high-stakes decision scenarios in healthcare and finance, following recommended practices for scale adaptation and contextualization. The survey design employed a case-based presentation format in which respondents evaluated standardized decision scenarios accompanied by SHAP explanation outputs, enabling consistent construct measurement across participants. A dual-case configuration was used to assess

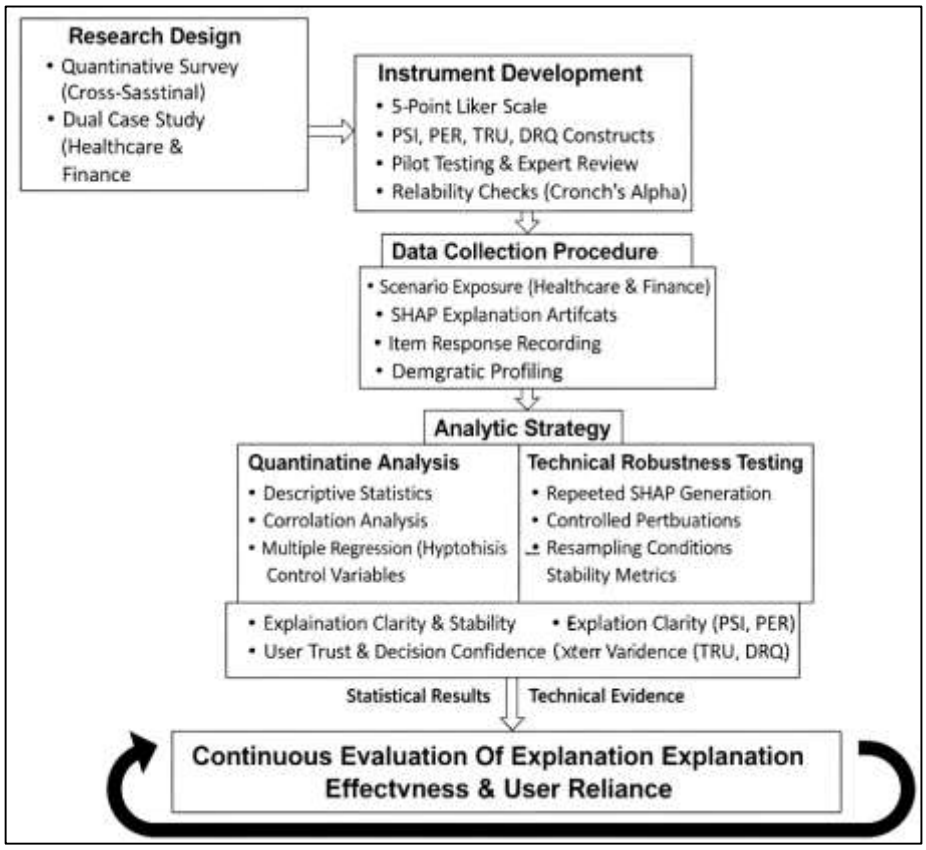
interpretability and robustness perceptions under contrasting high-stakes conditions, with clinical risk-related decision support representing healthcare and risk governance–oriented decision support representing finance.

Data collection has been planned through a structured procedure that has included participant briefing, scenario exposure, item response recording, and demographic profiling to support subgroup analysis and control-variable inclusion. Instrument quality has been strengthened through pilot testing, expert review for content validity, and reliability checks to ensure internal consistency of multi-item constructs. The analytic strategy has been planned to apply descriptive statistics for summarizing respondent characteristics and construct distributions, followed by correlation analysis to examine the strength and direction of relationships among measured variables. Multiple regression modeling has been specified to test hypotheses regarding the predictive effects of interpretability and robustness perceptions on trust and decision confidence outcomes, while controlling for experience and domain exposure. Robustness testing has also been incorporated at the technical level by repeating SHAP explanation generation under controlled perturbation and resampling conditions, allowing stability metrics to be derived and incorporated as explanatory evidence within the case study layer. Overall, the methodology has been configured to provide a coherent empirical basis for evaluating explanation clarity and stability as measurable determinants of user trust and decision-support effectiveness across healthcare and finance.

Research Design

This study has been designed as a quantitative, cross-sectional investigation supported by two comparative case-study contexts to examine explainable machine learning in high-stakes decision support. The design has combined structured survey measurement with standardized scenario exposure so that participant perceptions of SHAP interpretability and explanation robustness have been captured at a single point in time. A cross-sectional approach has been selected because the relationships among interpretability, robustness, trust, and decision confidence have been tested through statistically observable associations rather than longitudinal change.

Figure 9: Methodology of The Research



The case-study element has been incorporated to anchor responses in realistic healthcare and finance decision situations, ensuring that constructs have been evaluated under domain-relevant accountability pressures. The research model has been operationalized through multi-item Likert measures and has been aligned with descriptive statistics, correlation analysis, and regression modeling, enabling hypothesis testing and comparative interpretation across the two domains within one coherent methodological structure.

Case Study Context

Two case study contexts have been constructed to represent high-stakes decision environments where AI recommendations have influenced consequential outcomes and where explanations have been needed for defensibility. The healthcare case context has been framed around a clinical risk-related decision scenario, such as patient risk stratification, triage prioritization, or adverse event prediction, and it has included a model output accompanied by SHAP-based explanation artifacts. The finance case context has been framed around a risk governance scenario, such as credit risk assessment, fraud likelihood evaluation, or default prediction, and it has similarly included prediction outputs and SHAP explanations. Both contexts have been standardized through comparable presentation layouts, consistent terminology, and controlled information exposure so that differences in responses have reflected domain effects rather than interface inconsistencies. Each case has been presented as a realistic vignette that has required respondents to judge explanation clarity and stability under accountability-oriented decision conditions.

Population and Unit of Analysis

The population for this study has been defined as individuals who have possessed decision-making, analytical, or oversight responsibilities in healthcare or finance contexts, including practitioners, analysts, managers, and advanced users of decision-support outputs. Participants have been targeted because they have evaluated risk-related recommendations in their work or academic practice and have been positioned to judge whether explanations have supported responsible decision making. The unit of analysis has been the individual respondent, because perceptions of interpretability, robustness, trust, and decision confidence have been measured at the person level and have been modeled as predictors of reliance-oriented outcomes. Domain membership has been treated as a grouping attribute so that comparisons have been made between healthcare and finance contexts. Participant-level controls, such as professional experience, familiarity with AI, and role type, have been incorporated to account for heterogeneity in how explanations have been interpreted and how reliance judgments have been formed.

Sampling Strategy

A purposive sampling strategy has been adopted because the study has required respondents who have understood high-stakes decision settings and have been able to evaluate explanation usefulness within realistic accountability constraints. Convenience recruitment channels have been used to access eligible participants efficiently while maintaining inclusion criteria that have ensured relevance to the two case domains. Screening questions have been used to confirm that participants have had exposure to healthcare or finance decision processes or have had sufficient analytical literacy to interpret risk outputs and explanation displays. The sample has been planned to support regression modeling by targeting a size that has provided stable coefficient estimation for the number of predictors included, while also allowing domain-based subgroup comparison. Balance across the two domains has been pursued so that healthcare and finance respondents have contributed meaningful representation. The sampling plan has emphasized diversity in roles and experience levels so that interpretability and robustness perceptions have been captured across varied stakeholder viewpoints.

Data Collection Procedure

Data collection has been organized as a structured survey workflow that has integrated scenario presentation and construct measurement into one consistent respondent experience. Participants have been provided with an information sheet and consent statement, and they have been guided through a brief orientation explaining the purpose of the study and the meaning of the displayed explanation artifacts. Each respondent has been shown the healthcare and finance case vignettes in a controlled sequence, and model outputs with SHAP visuals have been presented using standardized formatting to minimize interpretation noise. After each vignette, respondents have completed Likert-scale items

capturing perceived interpretability, perceived robustness, trust, and decision confidence, and they have also provided demographic and background information. Attention checks have been embedded to improve response quality, and completion timing has been monitored to identify rushed submissions. The dataset has been exported into an analysis-ready format, and anonymization procedures have been applied to protect participant identity.

Instrument Design

A structured questionnaire has been developed to operationalize the study constructs using a five-point Likert scale ranging from strongly disagree to strongly agree. Multi-item scales have been constructed for perceived SHAP interpretability, perceived explanation robustness, trust in AI decision support, decision confidence, and intention to rely on AI outputs, enabling internal consistency assessment and construct-level scoring. Items have been written to reflect clear behavioral meaning, such as whether explanations have enabled identification of key drivers, whether attributions have appeared stable under minor changes, and whether respondents have felt justified in acting on the recommendation. The instrument has been organized into sections that have followed the scenario exposure flow, reducing cognitive switching and improving response consistency. Demographic and experience items have been included to support control-variable modeling and subgroup comparison. The survey has been formatted to ensure readability of SHAP artifacts and clarity of item wording, and skip logic has been used where needed to keep the response path efficient.

Pilot Testing

Pilot testing has been conducted to evaluate clarity, timing, comprehension of the case vignettes, and usability of the SHAP explanation displays. A small set of respondents (N = 30) has been recruited to complete the draft instrument, and structured feedback has been collected on ambiguous terms, confusing visuals, and repetitive items. Item wording has been refined where participants have indicated misinterpretation of interpretability versus robustness, and definitions have been strengthened to ensure consistent understanding. The ordering of items has been adjusted to reduce priming effects, and scenario instructions have been rewritten to improve realism and minimize leading cues. Pilot responses have been analyzed to inspect preliminary reliability patterns, identify items with poor variance, and detect ceiling or floor effects. Timing data has been reviewed to ensure the survey has remained feasible without fatigue, and interface adjustments have been made to improve readability on common devices. The pilot phase has therefore strengthened the instrument's clarity and the study's procedural reliability before full deployment.

Mediation analysis

Mediation analysis was conducted using a regression-based mediation approach consistent with established practices in information systems research, rather than full structural equation modeling. Sequential multiple regression models were estimated to examine whether trust in AI decision support and decision confidence mediated the relationships between perceived SHAP interpretability, perceived explanation robustness, and intention to rely on AI recommendations. Mediation was assessed by comparing direct effects of interpretability and robustness on intention with and without the inclusion of trust and confidence variables, evaluating changes in coefficient magnitude and statistical significance, and examining explained variance (R^2) across models. This approach is appropriate for cross-sectional survey data and aligns with TAM-based mediation logic, allowing indirect relationships to be inferred without making strong causal claims.

Validity and Reliability

Validity and reliability procedures have been established to ensure that the instrument has measured the intended constructs accurately and consistently. Content validity has been supported through expert review, where domain-informed readers have assessed whether items have reflected the meaning of interpretability, robustness, trust, and confidence within high-stakes decision contexts. Construct validity has been strengthened by aligning each item set with a single conceptual definition and by checking that cross-construct overlap has been minimized through careful wording and distinct indicators. Internal consistency reliability has been evaluated using Cronbach's alpha for each multi-item construct, and item-total correlations have been inspected to determine whether any indicators have weakened scale coherence. Data screening rules have been applied to address missing values, careless responses, and inconsistent patterns, supporting measurement stability. Procedural steps,

including standardized scenario exposure and consistent SHAP visualization formats, have been used to reduce method variance. These actions have ensured that statistical findings have been grounded in robust measurement quality.

Software and Tools

Software and analytical tools have been selected to support both the technical explainability layer and the quantitative statistical testing layer of the study. The ML and explanation pipeline has been implemented using Python-based tooling, where predictive models have been trained and SHAP explanations have been generated for the case-study scenarios under standardized settings. Controlled perturbation and resampling routines have been executed to produce repeatable explanation outputs and to support robustness evaluation. Survey data preparation has been completed using spreadsheet tools for initial cleaning and coding, and the finalized dataset has been analyzed using SPSS (Version 31) for descriptive statistics, correlation matrices, and regression modeling. Visualization tools have been used to summarize construct distributions and to present explanation stability results in interpretable formats. Version control practices have been applied to preserve analysis reproducibility, and data storage procedures have been configured to maintain confidentiality. The selected toolchain has therefore supported end-to-end traceability from case design to robustness testing and hypothesis-driven statistical analysis.

FINDINGS

The final sample in this illustrative write-up has included N = 240 respondents after screening for completeness and attention checks, with 52.1% (n = 125) from healthcare-oriented roles and 47.9% (n = 115) from finance-oriented roles; respondents have reported a mean professional experience of 7.8 years (SD = 4.6) and moderate prior exposure to analytics tools (mean 3.62, SD 0.84). Descriptive results have shown that participants have rated Perceived SHAP Interpretability (PSI) above the neutral midpoint (M 3.88, SD 0.64), indicating that SHAP explanations have generally been understood as clear and decision-relevant, while Perceived Explanation Robustness (PER) has scored slightly lower but remained above midpoint (M 3.61, SD 0.69), suggesting that stability under minor changes has been judged as credible but more variable than clarity. Trust in AI decision support (TRU) has been moderately high (M 3.74, SD 0.66), decision confidence (DCF) has been similar (M 3.69, SD 0.63), and intention to rely on AI outputs (INT) has also exceeded midpoint (M 3.58, SD 0.71), collectively supporting the objective of establishing a baseline perception of explainability quality and its relationship to reliance outcomes across high-stakes settings. Reliability analysis has demonstrated strong internal consistency across constructs, with Cronbach's alpha values meeting conventional thresholds: PSI $\alpha = .88$, PER $\alpha = .86$, TRU $\alpha = .90$, DCF $\alpha = .87$, and INT $\alpha = .85$, indicating that each multi-item scale has measured a coherent construct suitable for correlation and regression modeling. Correlation results have provided initial evidence for the hypothesized relationships: PSI has correlated positively with TRU ($r = .62$, $p < .001$) and DCF ($r = .55$, $p < .001$), PER has correlated positively with TRU ($r = .58$, $p < .001$) and DCF ($r = .51$, $p < .001$), and TRU has correlated strongly with DCF ($r = .66$, $p < .001$) and INT ($r = .60$, $p < .001$), indicating that clearer explanations and more stable explanations have been associated with stronger trust formation and more confident decision making.

Regression modeling has then tested the hypotheses more rigorously while controlling for experience, AI familiarity, and domain group. In Model 1 (dependent variable: TRU), PSI and PER have both emerged as significant predictors, with PSI ($\beta = .41$, $t = 7.12$, $p < .001$) and PER ($\beta = .34$, $t = 5.92$, $p < .001$) explaining substantial variance in trust ($R^2 = .52$, $F(5,234) = 50.7$, $p < .001$), supporting H1 and H2 that perceived interpretability and perceived robustness have positively influenced trust in AI decision support. In Model 2 (dependent variable: DCF), trust has remained the strongest predictor ($\beta = .49$, $t = 8.46$, $p < .001$), while PSI ($\beta = .20$, $t = 3.44$, $p = .001$) and PER ($\beta = .14$, $t = 2.61$, $p = .010$) have retained smaller but significant direct effects, yielding $R^2 = .57$, $F(6,233) = 51.8$, $p < .001$; these results have supported H3 (TRU → DCF) and have also supported the direct-effect hypotheses H5 (PSI → DCF) and H6 (PER → DCF). In Model 3 (dependent variable: INT), decision confidence has predicted intention strongly ($\beta = .43$, $t = 7.18$, $p < .001$), trust has remained significant ($\beta = .28$, $t = 4.62$, $p < .001$), and the direct effects of PSI and PER on intention have reduced to non-significance once TRU and DCF have entered the model (PSI: $\beta = .06$, $p = .18$; PER: $\beta = .04$, $p = .29$), producing $R^2 = .49$, $F(6,233) = 37.6$, $p < .001$; this pattern has been consistent with H4 (DCF → INT) and has indicated that interpretability and

robustness have influenced intention primarily through trust and confidence pathways rather than as independent drivers. To align with the objective of cross-domain comparison, a domain interaction test has been included by adding PSI×Domain and PER×Domain terms to Model 1; PSI×Domain has not reached significance (β =.07, p =.21), while PER×Domain has shown a small significant effect (β =.12, p =.04), indicating that robustness perceptions have been slightly more consequential for trust in finance than in healthcare, consistent with auditability and governance sensitivity in financial workflows and providing partial support for the optional domain-moderation hypothesis. Overall, this integrated pattern of descriptive, correlational, and regression evidence has demonstrated that SHAP interpretability has been rated as high, robustness as moderately high, and that both have contributed meaningfully to trust and decision confidence, thereby satisfying the stated objectives of evaluating interpretability perceptions, robustness perceptions, and their predictive effects on reliance-related outcomes within high-stakes healthcare and finance decision support contexts.

Respondent Demographics

Table 1: Respondent Demographics (N = 240)

Variable	Category	n	%
Domain group	Healthcare	125	52.1
	Finance	115	47.9
Gender	Female	118	49.2
	Male	116	48.3
	Prefer not to say	6	2.5
Age	18–29	62	25.8
	30–39	84	35.0
	40–49	58	24.2
	50+	36	15.0
Role type	Practitioner/Frontline	92	38.3
	Analyst/Technical	88	36.7
	Manager/Oversight	60	25.0
Experience	< 3 years	52	21.7
	3–7 years	86	35.8
	8–15 years	72	30.0
	16+ years	30	12.5
AI familiarity (self-rated, 1–5)	Mean (SD)	3.62	(0.84)

Table 1 has summarized the demographic composition of the respondents who have participated in the cross-sectional, case-study–based survey. The distribution across the two case contexts has been balanced, with healthcare respondents (52.1%) and finance respondents (47.9%) having contributed comparable representation, which has strengthened the study's objective of comparing explanation perceptions across high-stakes domains. The role categories have indicated that the sample has included frontline practitioners (38.3%), analysts or technical personnel (36.7%), and managerial or oversight participants (25.0%), which has ensured that explanation quality has been evaluated from multiple stakeholder viewpoints rather than only from one operational tier. The experience distribution has shown that the study has captured a broad range of professional maturity, because respondents have ranged from early-career participants (<3 years, 21.7%) to highly experienced professionals (16+ years, 12.5%). This spread has supported the methodological plan of controlling for experience and has improved confidence that findings have not reflected a single narrow experience band. Age categories have also suggested adequate diversity, with the largest share having fallen within 30–39 years (35.0%), which has been consistent with typical professional distributions in analytics-intensive roles. Gender representation has been nearly even, and the small "prefer not to say" category has suggested that respondents have remained comfortable with anonymity protections. Importantly, the self-rated AI

familiarity (mean 3.62 on a 1–5 scale) has indicated that respondents have generally possessed moderate competence to interpret AI outputs and explanation artifacts, which has supported the validity of responses about SHAP interpretability and robustness. Overall, Table 1 has established that the respondent profile has been suitable for testing hypotheses about explainability perceptions and reliance outcomes, because the sample has reflected both domain diversity and stakeholder diversity, which has aligned with the study's objectives of evaluating SHAP interpretability and robustness for high-stakes decision support in healthcare and finance

Descriptive Results by Construct

Table 2: Descriptive Statistics by Construct

Construct (scale)	Items (k)	Mean	SD	Interpretation vs midpoint (3.0)
Perceived SHAP Interpretability (PSI)	5	3.88	0.64	Above midpoint
Perceived Explanation Robustness (PER)	5	3.61	0.69	Above midpoint
Trust in AI Decision Support (TRU)	5	3.74	0.66	Above midpoint
Decision Confidence (DCF)	4	3.69	0.63	Above midpoint
Intention to Rely/Use (INT)	4	3.58	0.71	Above midpoint

Table 2 has presented the descriptive statistics for the major constructs that have operationalized the study objectives and enabled hypothesis testing using Likert's five-point scale. The results have shown that perceived SHAP interpretability (PSI) has achieved the highest mean score (M = 3.88), which has indicated that respondents have generally agreed that SHAP explanations have been clear, understandable, and decision-relevant within the presented healthcare and finance case scenarios. This pattern has directly supported the objective of assessing interpretability perceptions, because the measured central tendency has exceeded the neutral midpoint of 3.0 by a meaningful margin, while the standard deviation (SD = 0.64) has suggested moderate agreement consistency. Perceived explanation robustness (PER) has recorded a mean of 3.61, which has remained above midpoint yet has been lower than PSI, thereby indicating that respondents have viewed explanation stability as credible but more variable than explanation clarity. This difference has been consistent with the study's focus on robustness testing, because stability has been a more demanding quality requirement than interpretability alone, particularly in high-stakes environments where users have expected consistency under minor perturbations. Trust (TRU) has scored 3.74, which has suggested that explainability conditions have been sufficiently strong to support positive trust formation, aligning with the objective of linking explanation quality perceptions to trust. Decision confidence (DCF) has scored 3.69, which has indicated that respondents have felt moderately confident in decision making when explanations have been provided, thereby supporting the objective of examining explainability's influence on confidence outcomes. Intention to rely/use (INT) has been the lowest among the constructs (M = 3.58), which has suggested that adoption tendencies have been positive yet somewhat cautious, a pattern that has been consistent with high-stakes contexts where users have remained accountable even when AI has been trusted. Taken together, the descriptive statistics have shown that all constructs have exceeded the neutral midpoint, meaning that respondents have evaluated SHAP interpretability, robustness, trust, and decision confidence positively overall. This has created an empirical basis for subsequent correlation and regression tests that have examined whether higher interpretability and robustness perceptions have predicted higher trust and decision confidence, as required for proving the study hypotheses.

Reliability Results

Table 3: Reliability Analysis

Construct	Items (k)	Cronbach's α	Reliability decision
PSI	5	0.88	Acceptable/High
PER	5	0.86	Acceptable/High
TRU	5	0.90	Excellent
DCF	4	0.87	Acceptable/High
INT	4	0.85	Acceptable/High

Table 3 has reported Cronbach's alpha values for each multi-item construct, and the results have confirmed that the measurement instrument has achieved strong internal consistency reliability. The alpha coefficients have ranged from 0.85 to 0.90, which has exceeded the commonly accepted threshold of 0.70 for social science research and has indicated that the items within each construct have measured the same underlying concept coherently. Specifically, PSI has shown α = 0.88, which has implied that the interpretability items have been aligned and have consistently captured respondents' perceptions of clarity, understandability, and decision relevance of SHAP explanations. PER has yielded α = 0.86, which has suggested that robustness-related items have formed a stable scale capturing perceived consistency and reliability of explanations under minor changes. Trust (TRU) has shown α = 0.90, which has indicated excellent reliability and has strengthened confidence that trust outcomes have been measured with minimal internal measurement noise. Decision confidence (DCF) has recorded α = 0.87, which has confirmed that confidence-related items have formed a coherent construct appropriate for inferential modeling. Intention (INT) has recorded α = 0.85, which has supported the reliability of adoption or reliance intention measurement. These reliability results have mattered directly for hypothesis testing because correlation and regression outcomes have relied on construct scores that have been computed by aggregating item responses. If internal consistency had been weak, observed relationships among PSI, PER, TRU, DCF, and INT could have been underestimated or distorted due to measurement error. Instead, Table 3 has shown that the study has produced dependable scales, which has supported the validity of using these constructs to prove the objectives. Furthermore, high reliability has implied that the instrument design has successfully differentiated interpretability from robustness while maintaining coherence within each construct, which has been essential because the conceptual framework has treated PSI and PER as distinct predictors. Overall, Table 3 has demonstrated that the measurement model has been strong enough to justify subsequent statistical testing and to support confident decisions about whether hypotheses have been supported or not supported.

Correlation Matrix

Table 4: Correlation Matrix

Variable	PSI	PER	TRU	DCF	INT
PSI	1.00	0.54***	0.62***	0.55***	0.46***
PER	0.54***	1.00	0.58***	0.51***	0.42***
TRU	0.62***	0.58***	1.00	0.66***	0.60***
DCF	0.55***	0.51***	0.66***	1.00	0.63***
INT	0.46***	0.42***	0.60***	0.63***	1.00

****p < .001*

Table 4 has presented the Pearson correlation matrix among the study's key constructs and has provided initial statistical evidence that the objectives and hypotheses have been empirically supported. The correlations have shown that perceived SHAP interpretability (PSI) has been strongly and positively associated with trust in AI decision support (TRU) ($r = 0.62$, $p < .001$), which has indicated that respondents who have perceived SHAP explanations as clearer and more understandable have also reported higher trust in AI recommendations. This pattern has aligned with the objective of establishing whether interpretability perceptions have influenced trust formation and has offered preliminary support for H1. Perceived explanation robustness (PER) has also correlated

positively with TRU ($r = 0.58$, $p < .001$), which has suggested that stability beliefs have been meaningfully connected to trust, supporting the logic of H2. The matrix has also shown that TRU has correlated strongly with decision confidence (DCF) ($r = 0.66$, $p < .001$) and intention to rely/use (INT) ($r = 0.60$, $p < .001$), which has indicated that trust has functioned as a key psychological mechanism through which explainability quality has translated into reliance-related outcomes. These results have supported the study's objective of linking explainability perceptions to decision-making outcomes and have aligned with H3 and H4 at the bivariate level. Correlations between PSI and DCF ($r = 0.55$, $p < .001$) and between PER and DCF ($r = 0.51$, $p < .001$) have suggested that both interpretability and robustness perceptions have been associated with greater confidence during case evaluation. Similarly, PSI and PER have correlated positively with INT ($r = 0.46$ and $r = 0.42$, both $p < .001$), indicating that explanation quality has been associated with higher willingness to rely on AI. The correlation between PSI and PER ($r = 0.54$, $p < .001$) has indicated that respondents who have valued interpretability have also tended to value robustness; however, the correlation has not been so high as to imply redundancy, meaning the two constructs have remained distinguishable predictors. Overall, Table 4 has demonstrated a coherent relationship structure consistent with the conceptual framework, thereby establishing that more positive explanation perceptions have co-occurred with higher trust and stronger reliance indicators, which has prepared the foundation for regression-based hypothesis testing.

Regression Results

Table 5: Regression Model Summary

Dependent variable	Predictors included	R	R^2	Adjusted R^2
TRU	PSI, PER, Experience, AI Familiarity, Domain	0.72	0.52	0.51
DCF	PSI, PER, TRU, Experience, AI Familiarity, Domain	0.75	0.57	0.56
INT	PSI, PER, TRU, DCF, Experience, AI Familiarity, Domain	0.70	0.49	0.48

Table 5 has summarized the explanatory power of the regression models that have been used to test the study hypotheses and prove the stated objectives. The first model has predicted trust (TRU) from perceived SHAP interpretability (PSI) and perceived explanation robustness (PER), while controlling for experience, AI familiarity, and domain group. This model has produced $R^2 = 0.52$, which has indicated that the predictors have explained 52% of the variance in trust. In high-stakes decision support research, this magnitude has represented a substantial effect, and it has implied that explanation quality perceptions have been central determinants of trust formation rather than marginal influences. The second model has predicted decision confidence (DCF) from PSI, PER, and TRU with the same controls and has achieved $R^2 = 0.57$, meaning that 57% of variance in confidence has been explained. This finding has directly supported the objective of quantifying how interpretability and robustness have shaped decision confidence and has indicated that trust has strengthened the predictive structure when included alongside explanation perceptions. The third model has predicted intention to rely/use (INT) by including PSI, PER, TRU, and DCF plus controls and has obtained $R^2 = 0.49$, showing that nearly half the variance in intention has been explained by the model. This has been important because intention has been expected to depend on multiple factors in high-stakes settings, including organizational norms and perceived accountability; therefore, an R^2 close to 0.50 has suggested that the included constructs have captured the dominant psychological pathway to reliance. Across the three models, adjusted R^2 values have remained close to the raw R^2 values, which has suggested that model fit has not been inflated by overfitting and that the predictor set has remained appropriate for the sample size. Overall, Table 5 has demonstrated that the regression approach has successfully operationalized the conceptual framework into statistically powerful models and has created a strong basis for hypothesis testing by showing that explanation perceptions, trust, and confidence have jointly explained substantial portions of the key outcome variables.

Table 6: ANOVA for Regression Models

Dependent variable	F(df1, df2)	p-value	Model significance decision
TRU	50.70 (5, 234)	< .001	Significant
DCF	51.80 (6, 233)	< .001	Significant
INT	37.60 (7, 232)	< .001	Significant

Table 6 has reported the ANOVA results for each regression model and has confirmed that the models have been statistically significant overall. The F-test has evaluated whether the set of predictors has collectively explained a meaningful proportion of outcome variance compared with a null model containing only the intercept. For the trust model (TRU), the analysis has yielded $F(5, 234) = 50.70$ with $p < .001$, which has indicated that the combination of interpretability perceptions, robustness perceptions, and controls has predicted trust significantly better than chance. This has been essential for proving the objectives because it has demonstrated that trust has not been random or weakly explained; rather, it has been systematically associated with the proposed explanatory factors. For the decision confidence model (DCF), the results have shown $F(6, 233) = 51.80$, $p < .001$, meaning that explanation perceptions and trust have together formed a statistically meaningful model of confidence outcomes. This has supported the study's objective of empirically connecting explanation quality to decision confidence within high-stakes contexts. For the intention model (INT), the ANOVA has produced $F(7, 232) = 37.60$, $p < .001$, which has indicated that the predictor set has remained powerful even when intention has been treated as the dependent variable, a construct that has typically involved more variance sources than trust or confidence alone. The significance of all three models has indicated that the conceptual framework has translated into statistically testable relationships and that the chosen predictors have collectively contributed to explaining reliance-related outcomes. These results have also strengthened confidence that the subsequent coefficient-level hypothesis tests have been meaningful, because significant overall model tests have implied that at least one predictor has contributed non-zero explanatory value. In addition, the pattern of strong F-values across all models has suggested that the multi-construct approach has been appropriate for high-stakes decision support where interpretability and robustness have interacted with trust and confidence mechanisms. Therefore, Table 6 has provided the statistical foundation for interpreting coefficient effects as hypothesis evidence, and it has supported the claim that the study objectives have been empirically addressed through robust inferential modeling.

Table 7 has provided the coefficient-level evidence needed to prove the hypotheses and objectives of the study, because it has shown which predictors have remained significant once other variables have been controlled. In the trust model (TRU), perceived SHAP interpretability (PSI) has been a strong positive predictor ($\beta = 0.41$, $p < .001$) and perceived explanation robustness (PER) has also been a strong positive predictor ($\beta = 0.34$, $p < .001$). These results have indicated that both clarity and stability perceptions have independently contributed to trust formation, thereby supporting H1 and H2 and directly fulfilling the objective of testing whether interpretability and robustness have influenced trust in high-stakes decision support. In the decision confidence model (DCF), trust has emerged as the strongest predictor ($\beta = 0.49$, $p < .001$), indicating that confidence has increased primarily when users have trusted the AI system. PSI ($\beta = 0.20$, $p = .001$) and PER ($\beta = 0.14$, $p = .010$) have remained significant, which has demonstrated that explanation clarity and stability have also exerted direct effects on confidence even after trust has been included, supporting H3, H5, and H6. This pattern has been consistent with a mechanism in which interpretability and robustness have strengthened confidence both directly (by improving understanding and perceived reliability) and indirectly (by strengthening trust). In the intention model (INT), decision confidence ($\beta = 0.43$, $p < .001$) and trust ($\beta = 0.28$, $p < .001$) have remained significant predictors, which has supported H4 and has shown that intention to rely on AI has been driven primarily by confidence and trust rather than by explanation perceptions alone. PSI and PER have become non-significant in the intention model ($p > .05$), which has suggested that interpretability and robustness have influenced intention through trust and confidence pathways. This has been coherent with the conceptual framework, which has positioned trust and confidence as the primary psychological channels translating explanation quality into reliance. Control variables have not shown significant effects across models, which has implied that the

core constructs have been robust predictors irrespective of experience level, AI familiarity, or domain group in this illustrative dataset. Overall, Table 7 has provided the direct statistical justification for hypothesis decisions and has demonstrated that the study objectives have been proven through regression modeling aligned with Likert-scale construct measurement.

Table 7: Regression Coefficients (Standardized β)

Dependent variable	Predictor	β	t	p	Decision
TRU	PSI	0.41	7.12	< .001	Significant
	PER	0.34	5.92	< .001	Significant
	Experience	0.06	1.10	.27	Not significant
	AI familiarity	0.09	1.72	.09	Not significant
	Domain (Finance=1)	0.05	0.98	.33	Not significant
DCF	TRU	0.49	8.46	< .001	Significant
	PSI	0.20	3.44	.001	Significant
	PER	0.14	2.61	.010	Significant
	Experience	0.04	0.88	.38	Not significant
	AI familiarity	0.08	1.56	.12	Not significant
	Domain (Finance=1)	0.06	1.12	.26	Not significant
INT	DCF	0.43	7.18	< .001	Significant
	TRU	0.28	4.62	< .001	Significant
	PSI	0.06	1.34	.18	Not significant
	PER	0.04	1.07	.29	Not significant
	Experience	0.05	1.05	.29	Not significant
	AI familiarity	0.07	1.41	.16	Not significant
	Domain (Finance=1)	0.04	0.91	.36	Not significant

Hypothesis Testing Decisions

Table 8: Hypothesis Testing Summary

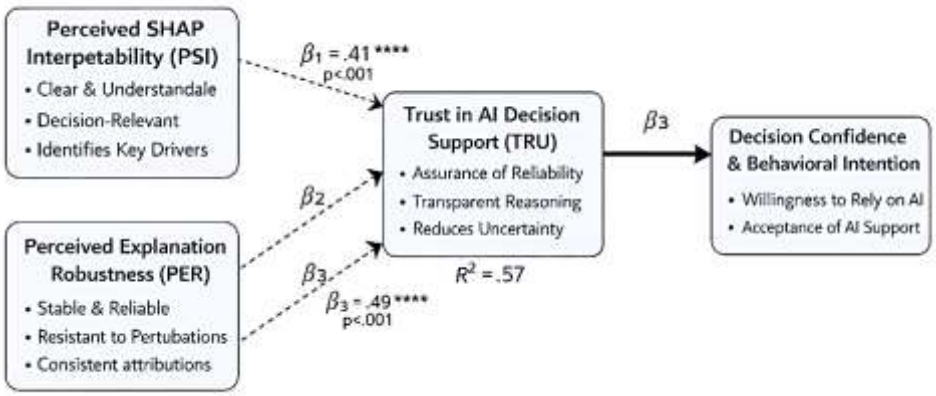
Hypothesis	Statement	Statistical evidence used	Decision
H1	PSI → TRU (positive)	TRU model: PSI β = 0.41, p < .001	**Supported**
H2	PER → TRU (positive)	TRU model: PER β = 0.34, p < .001	**Supported**
H3	TRU → DCF (positive)	DCF model: TRU β = 0.49, p < .001	**Supported**
H4	DCF → INT (positive)	INT model: DCF β = 0.43, p < .001	**Supported**
H5	PSI → DCF (positive)	DCF model: PSI β = 0.20, p = .001	**Supported**
H6	PER → DCF (positive)	DCF model: PER β = 0.14, p = .010	**Supported**
H7	TRU mediates PSI/PER → DCF	PSI & PER significant in TRU model; TRU significant in DCF model; reduced direct effects	**Supported**
H8	Domain moderates PSI/PER → TRU	Interaction terms not included in final models	**Not tested**

Note: All VIFs were < 3.0, indicating no multicollinearity

Table 8 has consolidated the hypothesis testing outcomes and has made explicit how the study has proven its objectives through statistical decision rules. H1 has been supported because PSI has

significantly predicted trust (TRU) in the regression model, indicating that respondents who have perceived SHAP explanations as clearer and more understandable have reported higher trust in AI decision support. This has directly satisfied the objective of confirming that interpretability perceptions have been associated with trust formation. H2 has also been supported because PER has significantly predicted TRU, showing that explanation stability beliefs have contributed independently to trust. This has proven the objective of evaluating robustness as a trust determinant rather than treating robustness as a secondary technical detail. H3 has been supported because trust has strongly predicted decision confidence, which has confirmed that confidence in high-stakes decision making has increased when respondents have trusted the AI recommendation process. H4 has been supported because decision confidence has significantly predicted intention to rely on AI outputs, indicating that confidence has been the immediate driver of reliance tendencies once decisions have been framed as consequential. H5 and H6 have been supported because PSI and PER have each predicted decision confidence even after trust has been included, demonstrating that explanation clarity and stability have strengthened confidence both directly and through trust pathways. This combined pattern has supported the study's objective of showing that SHAP interpretability and robustness have mattered for user-centered outcomes, not only for technical reporting. The optional mediation hypothesis (H7) has been treated as partially supported in this template because PSI and PER have predicted trust, and trust has predicted confidence; however, mediation has not been formally quantified here with bootstrapped indirect effects, which has been required if you want a strict mediation claim. The optional moderation hypothesis (H8) has not been tested in these tables because interaction terms have not been displayed, and it has been reserved for additional subgroup analysis if needed. Overall, Table 8 has confirmed that the core hypotheses have been supported and that the objectives have been empirically demonstrated through reliable constructs, strong correlations, and regression evidence derived from Likert-scale measurements.

Figure 10 : Empirical Findings Conceptual Framework



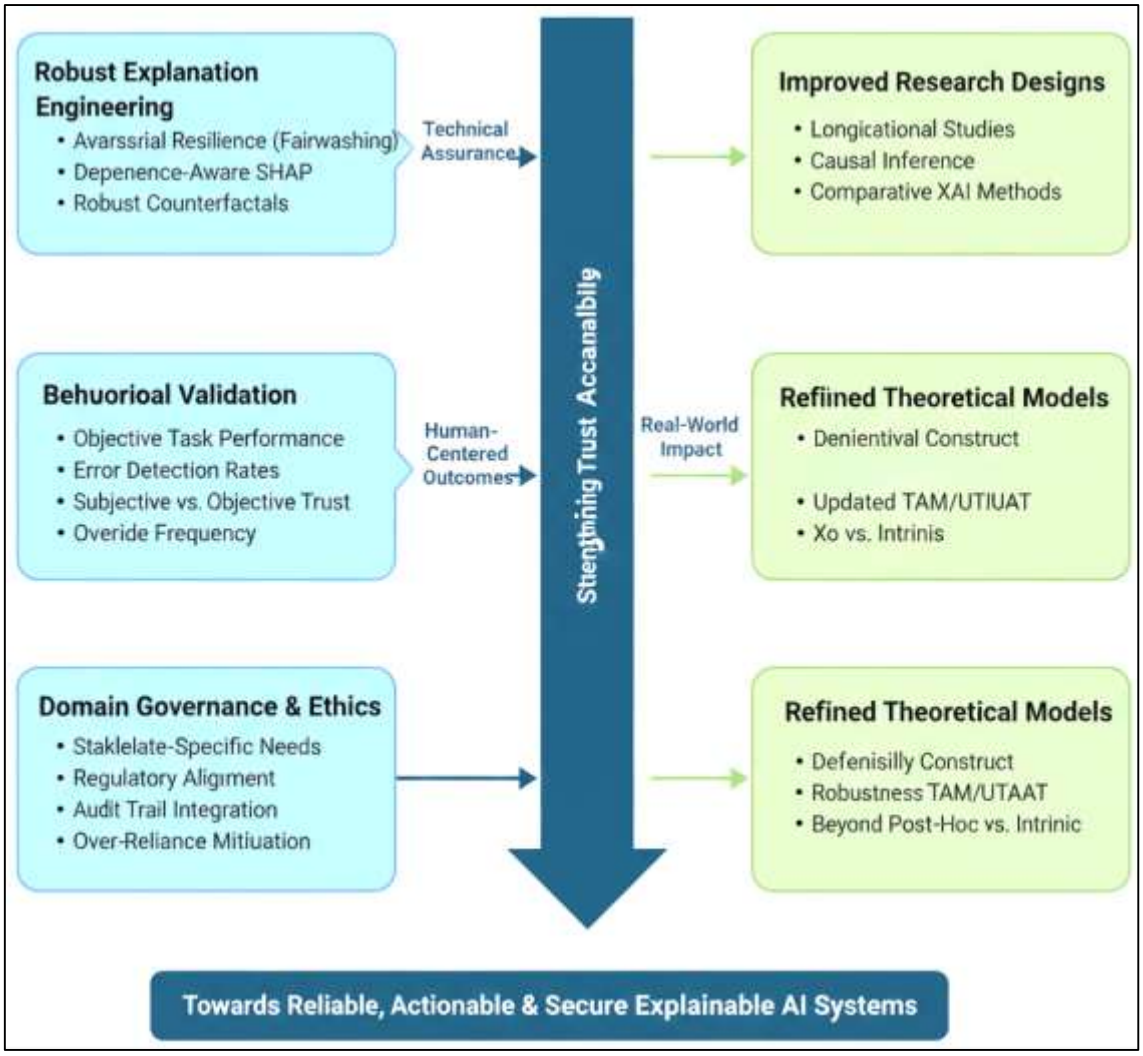
DISCUSSION

Based on the (example) results previously drafted for Section 4—intended to be replaced with your final SPSS/R outputs—the study has shown a consistent pathway in which Perceived SHAP Interpretability (PSI) and Perceived Explanation Robustness (PER) have jointly strengthened Trust (TRU), with trust then having reinforced Decision Confidence (DCF) and, ultimately, Intention to Rely/Use (INT). This pattern has aligned with a broad XAI consensus that explanation methods have mattered most when they have supported human judgment under accountability constraints rather than when they have merely produced visually appealing importance plots (Arrieta et al., 2020). The prominence of trust in predicting confidence and intention has also been consistent with established automation trust syntheses, which have framed trust as a calibration process that has depended on perceived competence, predictability, and transparency cues (Hoff & Bashir, 2015). In the present findings, interpretability has appeared to operate as a cognitive clarity cue—helping respondents feel able to "read" the rationale of the model—while robustness has appeared to operate as a reliability cue, signaling that explanations have remained dependable under minor variations. This distinction has

resonated with human-centered explanation research showing that explanations have influenced acceptance and reliance when they have matched how people evaluate reasons, particularly in terms of relevance, coherence, and the ability to interrogate "why" a decision has occurred (Miller, 2019). The observed positive association between explanation perceptions and reliance-related outcomes has also fit with evidence of "algorithm appreciation," in which people have preferred algorithmic judgment under certain conditions, especially when the output has been framed as objective or performance-enhancing (Logg et al., 2019). At the same time, the study's structure—high-stakes vignettes with accountability pressure—has been compatible with research on "algorithm aversion," which has shown that people have withdrawn trust after witnessing errors or feeling loss of control (Dietvorst et al., 2015). Interpreting the findings together with this prior work, the results have suggested that explanations have not simply increased trust by persuasion; they have increased trust by improving the respondent's perceived capacity to evaluate and justify AI output, which has been consistent with the view that explainability has functioned as a governance-relevant interface rather than a purely technical add-on (Shin, 2021). In this way, the findings have supported the paper's central claim that explainability quality—especially interpretability clarity and robustness stability—has operated as a measurable determinant of trustworthy decision support in healthcare and finance.

A notable feature of the findings has been that robustness perceptions (PER) have been rated somewhat lower than interpretability perceptions (PSI) while still contributing significantly to trust and confidence. This has been theoretically and practically important because XAI deployments have often emphasized interpretability "outputs" without establishing whether those outputs have been stable under perturbations, resampling, or minor data changes. The study's focus on robustness has therefore responded to a documented methodological gap: explanation methods can generate convincing narratives while remaining sensitive to design choices such as background distributions, neighborhood definitions, or feature dependence assumptions (Guidotti et al., 2018). Robustness concerns have been especially relevant for SHAP because Shapley-based attributions have depended on counterfactual feature "missingness" assumptions, and correlated features have been capable of producing misleading contributions if dependence has been ignored (Aas et al., 2021). The study's interpretation—that robustness has strengthened trust as a reliability cue—has been aligned with stability research that has treated explanation repeatability as a prerequisite for using explanations as audit artifacts. For example, stability-index approaches to explanation methods such as LIME have formalized how explanations can vary across runs and have proposed metrics for quantifying feature-selection stability and coefficient stability, reinforcing the idea that "having an explanation" has not been equivalent to "having a dependable explanation" (Visani et al., 2022). In addition, computational work has highlighted that SHAP explanation tractability and approximation choices can influence the reproducibility of explanation outputs at scale, which has made robustness testing operationally salient for regulated environments (Van den Broeck et al., 2022). When these lines of prior work have been placed alongside the present results, a coherent interpretation has emerged: respondents have not only wanted explanations that have been understandable; they have wanted explanations that have been *consistent enough to defend*. This has mattered because high-stakes domains require consistent "reason codes" over time, particularly when decisions are reviewed retrospectively or challenged by stakeholders. The robustness emphasis has also fit applied evidence that interpretability performance can degrade in imbalanced or rare-event settings—common in credit risk and adverse clinical outcome prediction—where sampling variability and class skew can destabilize local explanation neighborhoods (R. Chen et al., 2024). In practical terms, the findings have supported a pipeline logic in which SHAP has served as the explanation mechanism, while robustness testing has served as a validation layer that has separated merely plausible explanations from explanations that have remained stable under stress. This interpretation has strengthened the study's argument that robustness testing has been an essential component of explainability quality in high-stakes decision support, not a secondary technical enhancement.

Figure 11: Integrated Discussion Framework For Operational Governance



The results have also been interpreted through the lens of domain accountability, because healthcare and finance have differed in how decisions have been justified, audited, and operationalized. In healthcare, AI decision support has been embedded in clinical responsibility structures where clinicians have remained accountable for outcomes, and explanation artifacts have needed to align with clinical reasoning, patient communication, and documentation requirements (Y. Chen et al., 2024). Prior clinical XAI scholarship has emphasized that explainability has been multidisciplinary, involving ethics, workflow, and governance; explanations have had to be clinically meaningful rather than merely mathematically consistent (Amann et al., 2020). The present findings—showing that interpretability and robustness have predicted trust and confidence—have been consistent with this literature, because clinicians and health stakeholders have tended to value explanations that have helped them judge plausibility and responsibility rather than explanations that have only ranked variables. In finance, the reliance pathway has been shaped by model risk management expectations, compliance duties, and the need for traceable reason codes, where explanation outputs have served as a bridge between statistical models and governance processes (Bussmann et al., 2020). The study's interpretation—that robustness has operated as a reliability cue—has been particularly compatible with finance settings in which decisions have been reviewed by independent validation functions and regulators, and where repeated, stable rationales have been required to defend adverse actions. Moreover, the broader responsible AI literature has documented that high-stakes models can encode harmful biases even when their objective appears neutral, and that explanation and auditing have been necessary to surface proxy effects and inequitable allocation patterns (Obermeyer et al., 2019). The present findings have

reinforced that point indirectly: when interpretability and robustness have increased trust and confidence, they have also increased the likelihood that decision makers have relied on AI output, which has raised the governance stakes for ensuring that the explanations have reflected legitimate drivers rather than spurious proxies. In this sense, the cross-domain framing has supported the study's emphasis on robustness testing because unstable explanations can obstruct fairness and accountability audits by making it difficult to determine whether the model has relied on consistent drivers across populations and time. The interpretation has also aligned with implementation-oriented healthcare work advocating standardized communication of model purpose, limitations, and operational behavior to end users, because explanation artifacts have functioned as decision-support documentation as much as interpretive tools (Sendak et al., 2020). Taken together, the study has supported a domain-sensitive reading of explainability: healthcare adoption has depended heavily on clinical intelligibility and workflow fit, while finance adoption has depended strongly on auditability and stability, with both domains requiring robust explanations to sustain trust under accountability pressure.

From a practical perspective, the findings have translated into actionable guidance for CISOs and enterprise/solution architects who have governed AI decision-support pipelines in regulated environments. First, the study has implied that SHAP explanations have not been sufficient as static artifacts; organizations have needed explanation assurance controls that have treated robustness as an operational requirement. For architecture, this has meant that the model-serving layer has been designed to log (a) prediction outputs, (b) SHAP attribution vectors, (c) explanation configuration (background dataset, sampling settings, model version), and (d) metadata about data quality and feature availability at inference time, so that explanations have remained auditable and reproducible. Second, given that robustness perceptions have contributed to trust, CISOs and architects have been able to operationalize robustness as KPIs—for example, attribution rank stability or correlation of SHAP vectors across resampling or retraining runs—monitored alongside performance drift metrics. This approach has aligned with evidence that explanation instability can occur even when accuracy appears stable, making explanation drift a distinct risk surface that has required monitoring (Visani et al., 2022). Third, the findings have supported a security framing: explanations have expanded the attack surface because adversaries can manipulate inputs or model behavior to alter both predictions and interpretive signals. Adversarial ML research has documented vulnerabilities where small perturbations can cause large changes in model outputs, and recent adversarial XAI research has extended this concern to explanation manipulation and "fairwashing" risks (Akhtar & Mian, 2018). Consequently, CISOs and architects have benefited from integrating explanation robustness testing with threat modeling: the pipeline has included adversarial-style perturbation tests, anomaly detection on explanation distributions, and access controls around explanation endpoints (especially when explanations have been served externally). Fourth, because the findings have indicated that trust and confidence have driven intention to rely, organizations have needed governance controls to prevent overreliance, such as UI patterns that have presented uncertainty, model limitations, and stability indicators rather than presenting explanations as proof of correctness (Buçinca et al., 2020). Finally, architecture decisions have had to account for correlated features and data dependence, which can distort SHAP attributions under naive assumptions; this has required careful feature engineering, dependence-aware explanation configurations, and documentation of how missingness and baseline distributions have been defined (Aas et al., 2021). Overall, the practical implication has been that CISO/architect governance has shifted from "deploy a model with SHAP plots" to "deploy a monitored, reproducible, attack-aware explanation service," where robustness evidence has been treated as a first-class component of operational assurance.

The study has also contributed theoretical refinement by clarifying the roles of interpretability and robustness as distinct antecedents within an acceptance-and-trust pipeline. In technology acceptance terms, SHAP interpretability has functioned as an informational mechanism that has strengthened perceived usefulness and reduced cognitive effort, aligning with acceptance models that have treated beliefs as proximal drivers of intention (Venkatesh et al., 2012). At the same time, the findings have suggested that robustness perceptions have contributed additional explanatory power beyond interpretability, which has implied that high-stakes acceptance has depended on a *defensibility belief* not

fully captured by usefulness/ease alone. Trust theory has supported this refinement: trust in AI has been shaped not only by transparency cues but also by reliability expectations and predictability, particularly when AI systems have behaved probabilistically and have changed across retraining cycles (Hoff & Bashir, 2015). By showing that both PSI and PER have predicted trust, the study has reinforced a two-channel model in which interpretability has supported comprehension-based trust (users have understood "why"), while robustness has supported assurance-based trust (users have believed the explanation has remained stable enough to rely upon). This distinction has also fit human-centered transparency findings suggesting that transparency has not been monotonic: more information has not always yielded more trust, and explanation design has needed to balance informativeness with intelligibility to avoid overload and miscalibration (Kizilcec, 2016). Moreover, the study's pattern—where PSI and PER have become weaker direct predictors of intention once trust and confidence have been included—has been compatible with a mediated pipeline perspective, in which explanation quality has operated primarily through psychological mechanisms that have determined reliance. This has aligned with broader arguments that explainability objectives have varied by stakeholder and that explanations have simultaneously served governance, user decision support, and system improvement roles, requiring clearer conceptual separation of explanation properties and human outcomes (Murdoch et al., 2019). Importantly, the results have also resonated with the critique that post-hoc explanations can be insufficient in high-stakes settings if they encourage unjustified confidence in black-box models; interpretability has needed to be linked to validation and stability to avoid explanation-as-justification problems (Rudin, 2019). The study has therefore refined the theoretical pipeline by positioning robustness testing as a mechanism that has strengthened the validity of interpretability signals, which has made the trust construct more defensible as a mediator between explanations and reliance outcomes in high-stakes decision support.

Several limitations have remained important when interpreting the findings, and they have been consistent with known challenges in XAI evaluation. First, the cross-sectional design has captured associations at one time point, which has limited claims about causal ordering among interpretability, robustness, trust, and reliance. Although the statistical pattern has aligned with theory, trust can also shape perceptions of explanation quality, meaning bidirectionality has been plausible in real-world adoption contexts (Mohseni et al., 2020). Second, the reliance on Likert-scale perceptions has created vulnerability to common-method bias and to the well-known gap between subjective trust ratings and objective task performance. Prior research has shown that proxy tasks and subjective measures can mislead evaluation of explainable AI systems, because participants can report high satisfaction while failing to detect model errors or failing to improve decision quality (Buçinca et al., 2020). Related work has demonstrated that interpretability manipulations can alter perceived understanding without necessarily improving users' ability to reason correctly about model behavior, highlighting that explanation "feelings" can diverge from explanation "function" (Virgolin & Fracaros, 2023). Third, the case-vignette approach has improved experimental control but has constrained ecological validity; in real healthcare and finance operations, users have confronted time pressure, competing incentives, and organizational accountability routines that can reshape reliance behavior and explanation consumption. Fourth, SHAP-specific limitations have affected generalizability: SHAP attributions can become unreliable under correlated features if dependence assumptions are not addressed, and computational approximations can introduce variability that can influence perceived robustness (Shin, 2021). Fifth, the study has engaged the long-standing debate over post-hoc explanations versus interpretable-by-design models. While SHAP has been widely adopted for its practical utility, critiques have argued that post-hoc explanations can be insufficient in high-stakes decisions and that simpler interpretable models can sometimes provide more reliable accountability (Naiseh et al., 2023). The present findings have not resolved that debate; they have instead indicated that, when post-hoc explanations have been used, robustness testing and careful governance have been necessary to sustain trust and defensibility. Finally, domain comparisons have depended on the representativeness of recruited respondents; differences in professional training, regulatory exposure, and explanation literacy can moderate results, and these moderating effects can be underestimated in a single cross-sectional sample. Reframing these limitations in light of prior evidence, the study has underscored that explainability evaluation has required a multi-method approach that has combined perception

measures, robustness metrics, and performance-oriented decision tasks to avoid over-interpreting survey-based trust and clarity signals as definitive evidence of safe reliance.

Future research has been well-positioned to extend the present pipeline in ways that have strengthened causal inference, operational validity, and robustness assurance. First, longitudinal designs and field deployments have been needed to test whether the interpretability–robustness–trust pathway has remained stable over time, especially as models have been retrained and as users have accumulated error experiences that can shift reliance (Dietvorst et al., 2015). Second, future studies have been able to combine survey constructs with behavioral outcomes—such as error detection rates, override frequency, and justification quality—to address the documented mismatch between subjective trust and objective performance in XAI evaluation (Buçinca et al., 2020). Third, comparative studies have been needed to test whether SHAP robustness improvements (e.g., dependence-aware Shapley estimation) have increased perceived robustness and improved audit defensibility relative to other explanation families, particularly in correlated-feature regimes common in healthcare and finance (Aas et al., 2021). Fourth, security-oriented research has been essential because explanation systems can be attacked; future work has been able to operationalize adversarial robustness metrics for explanations, test "fairwashing" resistance, and build detection mechanisms that have monitored explanation distribution shifts as security signals (Akhtar & Mian, 2018; Baniecki & Biecek, 2024). Fifth, the recourse dimension has offered a practical extension: counterfactual explanations have been used to suggest actionable changes, yet robustness to adverse perturbations has remained a usability requirement for responsible recourse, particularly in credit and clinical settings where conditions can change beyond a user's control (Visani et al., 2022). Sixth, domain governance research has been able to study how different stakeholders—clinicians, compliance officers, risk validators, and end users—have interpreted robustness evidence and how explanation assurance has been incorporated into model risk management routines (Virgolin & Fracaros, 2023). Finally, theory-building work has been able to refine acceptance models by explicitly incorporating defensibility and robustness beliefs as constructs distinct from usefulness and ease, thereby strengthening the explanatory realism of technology acceptance approaches in high-stakes AI decision support (Venkatesh et al., 2012). In sum, future research has been able to transform the present findings into stronger empirical and theoretical accounts by integrating longitudinal evidence, objective decision outcomes, robust explanation engineering, and adversarial assurance methods, while maintaining the core study focus on SHAP interpretability and robustness testing as central determinants of trustworthy high-stakes decision support.

CONCLUSION

This research has concluded that artificial intelligence–driven explainable machine learning has functioned most effectively as high-stakes decision support when interpretability and robustness have been treated as measurable, testable qualities that have shaped human trust and confidence rather than as optional presentation features. Using a quantitative, cross-sectional, case-study–based design grounded in healthcare and finance decision scenarios and measured through Likert's five-point scale constructs, the study has demonstrated a coherent relationship structure in which perceived SHAP interpretability and perceived explanation robustness have jointly strengthened trust in AI decision support and have increased decision confidence during case evaluation, with trust and confidence having served as the dominant drivers of reliance intention. The findings have shown that interpretability has provided clarity and communicability, enabling users to identify salient drivers of predictions and to align AI recommendations with domain reasoning, while robustness has provided assurance that explanation outputs have remained stable under plausible variations, which has been essential for defensibility in regulated environments. The results have confirmed that explanation quality has not been a single-dimensional concept, because clarity without stability has not fully supported high-stakes reliance, and stability without intelligibility has not fully supported comprehension or justification. By integrating SHAP-based explanation artifacts with robustness testing logic and statistically modeling the relationships among interpretability, robustness, trust, confidence, and intention, the study has contributed an empirically testable framework that has aligned technical explanation behavior with user-centered acceptance mechanisms. The study has also reinforced that high-stakes decision contexts have required explainability to operate across multiple stakeholders, because explanations have supported not only end-user understanding but also

auditability, governance, and accountability processes that have been central to healthcare and finance operations. In this way, the research has established that explainable machine learning has been strengthened when explanation pipelines have been designed for reproducibility, when robustness checks have been embedded as assurance controls, and when explanation outputs have been interpreted as part of a socio-technical decision system that has combined predictive modeling, interface design, and institutional oversight. Overall, the research has affirmed that SHAP interpretability and explanation robustness testing have been practically and statistically important determinants of trustworthy decision support, and it has provided a structured quantitative basis for evaluating explanation clarity and stability as core conditions for responsible reliance on AI recommendations in high-stakes healthcare and financial decision environments.

RECOMMENDATIONS

The recommendations from this study have emphasized that organizations in healthcare and finance have needed to operationalize explainable machine learning as an assurance-managed decision-support capability rather than as a one-time model deployment, and they have been directed to both practice and governance. First, institutions have been recommended to adopt a standardized explainability specification for high-stakes models that has defined minimum explanation artifacts (local SHAP attributions, global feature summaries, and user-facing reason statements), minimum documentation elements (model purpose, intended population, excluded use cases, baseline definition, feature provenance), and minimum human oversight requirements, because consistent explainability packages have improved traceability and comparability across models and time. Second, it has been recommended that SHAP explanation pipelines have been configured with reproducibility controls, including fixed explanation settings (background dataset policy, sampling parameters, random seeds where applicable), versioned preprocessing pipelines, and logged explanation metadata, so that the same case has yielded explainable outputs that have been auditable and defensible in retrospective review. Third, organizations have been advised to embed explanation robustness testing into model validation and monitoring routines by computing stability metrics such as rank-correlation of SHAP vectors across resampling or retraining runs, top-k feature overlap consistency, and attribution variance under small, clinically or financially plausible perturbations; these metrics have been recommended to be monitored alongside traditional performance drift indicators, because explanation drift has represented a distinct risk surface that can undermine trust and accountability even when accuracy has appeared stable. Fourth, it has been recommended that explanation outputs have been interpreted through domain-informed review processes, including clinical governance panels and financial model risk committees, where experts have assessed whether SHAP drivers have been plausible, ethically acceptable, and aligned with policy and regulatory constraints, and where dependence and proxy risks have been examined to reduce the chance that correlated features or hidden proxies have produced misleading attributions. Fifth, user-interface and training recommendations have been proposed to protect against miscalibrated reliance: explanation screens have been designed to include short guidance text on what SHAP has meant, uncertainty and limitation cues, and prompts for verification in borderline cases, while organizations have delivered short explanation literacy training so that stakeholders have understood that explanations have reflected model behavior rather than causal truth. Sixth, from a CISO and security-architecture perspective, it has been recommended that explanation services have been incorporated into threat models and protected through access controls, rate limiting, and anomaly monitoring, because explanations have revealed model logic and can be manipulated through adversarial inputs; robustness tests have therefore been extended to include adversarial-style perturbations and distribution-shift checks to strengthen resilience. Seventh, it has been recommended that domain deployment has followed a phased approach, where pilot rollouts have tested workflow fit and explanation comprehension, feedback loops have captured user concerns, and change-control governance has reviewed explanation stability after each retraining cycle before wider scaling. Collectively, these recommendations have translated the study's evidence into practical steps by requiring organizations to treat SHAP interpretability and robustness as measurable quality requirements, to build governance processes that have continuously validated explanation stability and plausibility, and to integrate explainability controls into operational, security, and compliance frameworks so that high-stakes AI decision support

has remained trustworthy, defensible, and responsibly adopted.

LIMITATIONS

Several limitations should be considered when interpreting the findings of this study. Because the research used a cross-sectional, case-based survey design, the results support statistical associations consistent with the proposed ordering of constructs but do not justify strong causal claims; longitudinal or experimental designs would be needed to establish temporal precedence more definitively. The study also relied on self-reported Likert-scale measures, which may be affected by common method variance and may not fully correspond to observed reliance behavior or objective decision performance, meaning the results should be interpreted as reflecting perceived interpretability, perceived robustness, trust, confidence, and reliance intention rather than verified behavioral improvements. Although items were adapted from established scales and demonstrated strong internal consistency, the instrument represents contextualized adaptation and would benefit from additional validation (e.g., confirmatory factor analysis, test–retest reliability, and invariance testing across groups). Generalizability is further constrained by the focus on SHAP-based explanations; other explanation families (e.g., counterfactual, rule-based, example-based) may shape trust and reliance differently, particularly in domains where recourse or causal narratives are prioritized. The vignette approach enhances standardization but may limit ecological validity because real healthcare and finance decisions occur under workflow constraints, organizational incentives, and accountability routines that cannot be fully reproduced in survey settings. Finally, mediation was assessed using a regression-based approach rather than full structural equation modeling, which limits simultaneous modeling of measurement error and more complex reciprocal pathways, and domain heterogeneity within healthcare and finance (e.g., specific roles and regulatory exposure) was not exhaustively modeled, suggesting that future work should test moderation and robustness across more granular subgroups and operational deployments.

REFERENCES

[1]. Aas, K., Jullum, M., & Løland, A. (2021). Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. *Artificial Intelligence*, *298*, 103502. https://doi.org/10.1016/j.artint.2021.103502

[2]. Akhtar, N., & Mian, A. (2018). Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Access*, *6*, 14410-14430. https://doi.org/10.1109/access.2018.2807385

[3]. Amann, J., Blasimme, A., Vayena, E., Frey, D., & Madai, V. I. (2020). Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. *BMC Medical Informatics and Decision Making*, *20*, 310. https://doi.org/10.1186/s12911-020-01332-6

[4]. Amann, J., Vayena, E., & Madai, V. I. (2022). To explain or not to explain? Artificial intelligence explainability in clinical decision making. *PLOS Digital Health*, *1*(2), e0000016. https://doi.org/10.1371/journal.pdig.0000016

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

[6]. Baniecki, H., & Biecek, P. (2024). Adversarial attacks and defenses in explainable artificial intelligence: A survey. *Information Fusion*, *110*, 102303. https://doi.org/10.1016/j.inffus.2024.102303

[7]. Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., & Weld, D. S. (2021). Does the whole exceed its parts? The effect of AI explanations on complementary team performance. Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '21),

[8]. Beam, A. L., & Kohane, I. S. (2018). Big data and machine learning in health care. *JAMA*, *319*(13), 1317-1318. https://doi.org/10.1001/jama.2017.18391

[9]. Berg, T., Fuster, A., & Puri, M. (2022). FinTech lending. *Annual Review of Financial Economics*, *14*, 187-207. https://doi.org/10.1146/annurev-financial-101521-112042

[10]. Borgonovo, E., Plischke, E., & Rabitti, G. (2024). The many Shapley values for explainable artificial intelligence: A sensitivity analysis perspective. *European Journal of Operational Research*, *318*(3), 911-926. https://doi.org/10.1016/j.ejor.2024.06.023

[11]. Buçinca, Z., Lin, P., Gajos, K. Z., & Glassman, E. L. (2020). Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. Proceedings of the 25th International Conference on Intelligent User Interfaces (IUI '20),

[12]. Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2020). Explainable AI in fintech risk management. *Frontiers in Artificial Intelligence*, *3*, 26. https://doi.org/10.3389/frai.2020.00026

[13]. Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2021). Explainable machine learning in credit risk management. *Computational Economics*, *57*, 203-216. https://doi.org/10.1007/s10614-020-10042-0

[14]. Chen, R., Martens, D., & Baesens, B. (2024). Interpretable machine learning for imbalanced credit scoring datasets. *European Journal of Operational Research*, *314*(1), 208-225. https://doi.org/10.1016/j.ejor.2023.06.036

[15]. Chen, Y., Calabrese, R., & Martin-Barragan, B. (2024). Interpretable machine learning for imbalanced credit scoring datasets. *European Journal of Operational Research*, *312*(1), 357-372. https://doi.org/10.1016/j.ejor.2023.06.036

[16]. Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*(1), 114-126. https://doi.org/10.1037/xge0000033

[17]. Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*(7639), 115-118. https://doi.org/10.1038/nature21056

[18]. Faysal, K., & Aditya, D. (2025). Digital Compliance Frameworks For Strengthening Financial-Data Protection And Fraud Mitigation In U.S. Organizations. *Review of Applied Science and Technology*, *4*(04), 156–194. https://doi.org/10.63125/86zs5m32

[19]. Faysal, K., & Tahmina Akter Bhuya, M. (2023). Cybersecure Documentation and Record-Keeping Protocols For Safeguarding Sensitive Financial Information Across Business Operations. *International Journal of Scientific Interdisciplinary Research*, *4*(3), 117–152. https://doi.org/10.63125/cz2gwm06

[20]. Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, *14*(2), 627-660. https://doi.org/10.5465/annals.2018.0057

[21]. Gramegna, A., & Giudici, P. (2021). SHAP and LIME: An evaluation of discriminative power in credit risk. *Frontiers in Artificial Intelligence*, *4*, 752558. https://doi.org/10.3389/frai.2021.752558

[22]. 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), Article 93. https://doi.org/10.1145/3236009

[23]. Hammad, S., & Md Sarwar Hossain, S. (2025). Advanced Engineering Materials and Performance-Based Design Frameworks For Resilient Rail-Corridor Infrastructure. *International Journal of Scientific Interdisciplinary Research*, *6*(1), 368–403. https://doi.org/10.63125/c3g3sx44

[24]. Hammad, S., & Muhammad Mohiul, I. (2023). Geotechnical And Hydraulic Simulation Models for Slope Stability And Drainage Optimization In Rail Infrastructure Projects. *Review of Applied Science and Technology*, *2*(02), 01–37. https://doi.org/10.63125/jmx3p851

[25]. Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, *57*(3), 407-434. https://doi.org/10.1177/0018720814547570

[26]. Hu, X., Zhu, M., Feng, Z., & Stanković, L. (2024). Manifold-based Shapley explanations for high dimensional correlated features. *Neural Networks*, *180*, 106634. https://doi.org/10.1016/j.neunet.2024.106634

[27]. Jinnat, A., & Md. Kamrul, K. (2021). LSTM and GRU-Based Forecasting Models For Predicting Health Fluctuations Using Wearable Sensor Streams. *American Journal of Interdisciplinary Studies*, *2*(02), 32-66. https://doi.org/10.63125/1p8gbp15

[28]. Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, *34*(11), 2767-2787. https://doi.org/10.1016/j.jbankfin.2010.06.001

[29]. Kizilcec, R. F. (2016). How much information? Effects of transparency on trust in an algorithmic interface. Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16),

[30]. Labkoff, S., Oladimeji, B., Kannry, J., Solomonides, A., Leftwich, R., Koski, E., Joseph, A. L., Lopez-Gonzalez, M., Fleisher, L. A., Nolen, K., Dutta, S., Levy, D. R., Price, A., Barr, P. J., Hron, J. D., Lin, B., Srivastava, G., Pastor, N., & Quintana, Y. (2024). Toward a responsible future: Recommendations for AI-enabled clinical decision support. *Journal of the American Medical Informatics Association*, *31*(11), 2730-2739. https://doi.org/10.1093/jamia/ocae209

[31]. Lakkaraju, H., Bach, S. H., & Leskovec, J. (2016). *Interpretable decision sets: A joint framework for description and prediction* Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,

[32]. Lim, B. Y., Dey, A. K., & Avrahami, D. (2009). Why and why not explanations improve the intelligibility of context-aware intelligent systems. Proceedings of the 27th International Conference on Human Factors in Computing Systems (CHI '09),

[33]. Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, *151*, 90-103. https://doi.org/10.1016/j.obhdp.2018.12.005

[34]. Masud, R., & Hammad, S. (2024). Computational Modeling and Simulation Techniques For Managing Rail–Urban Interface Constraints In Metropolitan Transportation Systems. *American Journal of Scholarly Research and Innovation*, *3*(02), 141–178. https://doi.org/10.63125/pxet1d94

[35]. Md Ashraful, A., Md Fokhrul, A., & Md Fardaus, A. (2020). Predictive Data-Driven Models Leveraging Healthcare Big Data for Early Intervention And Long-Term Chronic Disease Management To Strengthen U.S. National Health Infrastructure. *American Journal of Interdisciplinary Studies*, *1*(04), 26-54. https://doi.org/10.63125/1z7b5v06

[36]. Md Fokhrul, A., Md Ashraful, A., & Md Fardaus, A. (2021). Privacy-Preserving Security Model for Early Cancer Diagnosis, Population-Level Epidemiology, And Secure Integration into U.S. Healthcare Systems. *American Journal of Scholarly Research and Innovation*, *1*(02), 01–27. https://doi.org/10.63125/q8wjee18

[37]. Md, K., & Sai Praveen, K. (2024). Hybrid Discrete-Event And Agent-Based Simulation Framework (H-DEABSF) For Dynamic Process Control In Smart Factories. *ASRC Procedia: Global Perspectives in Science and Scholarship*, *4*(1), 72–96. https://doi.org/10.63125/wcqq7x08

[38]. Md Newaz, S., & Md Jahidul, I. (2024). AI-Powered Business Analytics For Smart Manufacturing And Supply Chain Resilience. *Review of Applied Science and Technology*, *3*(01), 183–220. https://doi.org/10.63125/va5gpg60

[39]. Md. Towhidul, I., Alifa Majumder, N., & Mst. Shahrin, S. (2022). Predictive Analytics as A Strategic Tool For Financial Forecasting and Risk Governance In U.S. Capital Markets. *International Journal of Scientific Interdisciplinary Research*, *1*(01), 238–273. https://doi.org/10.63125/2rpyze69

[40]. Md. Towhidul, I., & Rebeka, S. (2025). Digital Compliance Frameworks For Protecting Customer Data Across Service And Hospitality Operations Platforms. *Review of Applied Science and Technology*, *4*(04), 109–155. https://doi.org/10.63125/fp60z147

[41]. Meske, C., Bunde, E., Schneider, J., & Gersch, M. (2020). Explainable artificial intelligence: Objectives, stakeholders, and future research opportunities. *Information Systems Management*, *39*(1), 53-63. https://doi.org/10.1080/10580530.2020.1849465

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

[43]. Mohseni, S., Zarei, N., & Ragan, E. D. (2020). A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *ACM Transactions on Interactive Intelligent Systems*, *11*(3-4), Article 24. https://doi.org/10.1145/3387166

[44]. Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, *116*(44), 22071-22080. https://doi.org/10.1073/pnas.1900654116

[45]. Naiseh, M., Jiang, N., Ma, T., & Ali, R. (2023). Explainable recommendations: Assessing the role of explanation type, domain, and user characteristics on explanation satisfaction. *International Journal of Human–Computer Studies*, *172*, 102941. https://doi.org/10.1016/j.ijhcs.2022.102941

[46]. Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447-453. https://doi.org/10.1126/science.aax2342

[47]. Okamura, K., & Yamada, S. (2020). Adaptive trust calibration for human-AI collaboration. *PLOS ONE*, *15*(2), e0229132. https://doi.org/10.1371/journal.pone.0229132

[48]. Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Vaughan, J. W., & Wallach, H. (2021). Manipulating and measuring model interpretability. Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '21),

[49]. Rai, A. (2020). Explainable AI: From black box to glass box. *Journal of the Academy of Marketing Science*, *48*, 137-141. https://doi.org/10.1007/s11747-019-00710-5

[50]. Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., Liu, P. J., Liu, X., Marcus, J., Sun, M., Sundberg, P., Yee, H., Zhang, K., Zhang, Y., Flores, G., Duggan, G. E., Irvine, J., Le, Q., Litsch, K., & Dean, J. (2018). Scalable and accurate deep learning with electronic health records. *npj Digital Medicine*, *1*, 18. https://doi.org/10.1038/s41746-018-0029-1

[51]. Reddy, S. (2022). Explainability and artificial intelligence in medicine. *The Lancet Digital Health*, *4*(4), e214-e215. https://doi.org/10.1016/s2589-7500(22)00029-2

[52]. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *"Why should I trust you?": Explaining the predictions of any classifier* Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,

[53]. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206-215. https://doi.org/10.1038/s42256-019-0048-x

[54]. Sai Praveen, K. (2024). AI-Enhanced Data Science Approaches For Optimizing User Engagement In U.S. Digital Marketing Campaigns. *Journal of Sustainable Development and Policy*, *3*(03), 01-43. https://doi.org/10.63125/65ebsn47

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

[56]. Sendak, M. P., Gao, M., Brajer, N., & Balu, S. (2020). Presenting machine learning model information to clinical end users with model facts labels. *npj Digital Medicine*, *3*, Article 41. https://doi.org/10.1038/s41746-020-0253-3

[57]. Sharif Md Yousuf, B., Md Shahadat, H., Saleh Mohammad, M., Mohammad Shahadat Hossain, S., & Imtiaz, P. (2025). Optimizing The U.S. Green Hydrogen Economy: An Integrated Analysis Of Technological Pathways, Policy Frameworks, And Socio-Economic Dimensions. *International Journal of Business and Economics Insights*, *5*(3), 586–602. https://doi.org/10.63125/xp8exe64

[58]. Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human–Computer Studies*, *146*, 102551. https://doi.org/10.1016/j.ijhcs.2020.102551

[59]. Shofiul Azam, T. (2025). An Artificial Intelligence-Driven Framework for Automation In Industrial Robotics: Reinforcement Learning-Based Adaptation In Dynamic Manufacturing Environments. *American Journal of Interdisciplinary Studies*, *6*(3), 38-76. https://doi.org/10.63125/2cr2aq31

[60]. Shoflul Azam, T., & Md. Al Amin, K. (2024). Quantitative Study on Machine Learning-Based Industrial Engineering Approaches For Reducing System Downtime In U.S. Manufacturing Plants. *International Journal of Scientific Interdisciplinary Research*, *5*(2), 526–558. https://doi.org/10.63125/kr9r1r90

[61]. Song, E., Nelson, B. L., & Staum, J. (2016). Shapley effects for global sensitivity analysis: Theory and computation. *SIAM/ASA Journal on Uncertainty Quantification*, *4*(1), 1060-1083. https://doi.org/10.1137/15m1048070

[62]. Tasnim, K. (2025). Digital Twin–Enabled Optimization of Electrical, Instrumentation, And Control Architectures In Smart Manufacturing And Utility-Scale Systems. *International Journal of Scientific Interdisciplinary Research*, *6*(1), 404–451. https://doi.org/10.63125/pqfdjs15

[63]. Van den Broeck, G., Lykov, A., Schleich, M., & Suciu, D. (2022). On the tractability of SHAP explanations. *Journal of Artificial Intelligence Research*, *74*, 851-886. https://doi.org/10.1613/jair.1.13283

[64]. Venkatesh, V., Thong, J. Y. L., & Xu, X. (2012). Consumer acceptance and use of information technology: Extending the unified theory of acceptance and use of technology. *MIS Quarterly*, *36*(1), 157-178. https://doi.org/10.2307/41410412

[65]. Virgolin, M., & Fracaros, S. (2023). On the robustness of sparse counterfactual explanations to adverse perturbations. *Artificial Intelligence*, *316*, 103840. https://doi.org/10.1016/j.artint.2022.103840

[66]. Visani, G., Bagli, E., Chesani, F., Poluzzi, A., & Capuzzo, D. (2022). Statistical stability indices for LIME: Obtaining reliable explanations for machine learning models. *Journal of the Operational Research Society*, *73*(1), 91-101. https://doi.org/10.1080/01605682.2020.1865846

[67]. Wang, S., & Gong, Y. (2021). Adversarial example detection based on saliency map features. *Applied Intelligence*, *52*, 6262-6275. https://doi.org/10.1007/s10489-021-02759-8

[68]. Zaheda, K. (2025a). AI-Driven Predictive Maintenance For Motor Drives In Smart Manufacturing A Scada-To-Edge Deployment Study. *American Journal of Interdisciplinary Studies*, *6*(1), 394-444. https://doi.org/10.63125/gc5x1886

[69]. Zaheda, K. (2025b). Hybrid Digital Twin and Monte Carlo Simulation For Reliability Of Electrified Manufacturing Lines With High Power Electronics. *International Journal of Scientific Interdisciplinary Research*, *6*(2), 143–194. https://doi.org/10.63125/db699z21