American Journal of Interdisciplinary Studies

Volume: 4; **Issue:** 3
Pages: 96-135
eISSN: 3067-5146

EXPLAINABLE ARTIFICIAL INTELLIGENCE (XAI) APPROACHES FOR CYBER RISK ASSESSMENT IN FINANCIAL SERVICES

Saba Ashfaq[1]; Tonoy Kanti Chowdhury[2];

[1]. *MS IT - Software Design and Management: Washington University of Science and Technology, USA; Email: sabarashfaq01@gmail.com*

[2]. *Master of Science in Information Technology, Washington University of Science and Technology, USA; Email: chowdhurytonoy93@gmail.com*

Doi: 10.63125/3gjcb322
Received: *16 June 2023;* ***Revised:*** *27 July 2023;* ***Accepted:*** *22 August 2023;* ***Published:*** *28 September 2023*

Abstract

This study investigated the quantitative influence of explainable artificial intelligence on cyber risk assessment within financial systems by evaluating how interpretability metrics contributed to model performance, alert clarity, and operational decision outcomes. Using three large-scale financial cybersecurity datasets consisting of 1.2 million fraud records, 4.8 million network intrusion events, and 930,000 authentication logs, the study analyzed both interpretable and non-interpretable machine learning models. Detection performance metrics, including precision, recall, F1-score, and AUC, were examined alongside interpretability measures such as fidelity, stability, and explanation complexity. Results showed that fidelity demonstrated strong positive correlations with performance metrics, ranging from r = .69 to r = .76 across datasets, while stability showed moderate to strong correlations (r = .64 to r = .72). Explanation complexity exhibited negative correlations with detection performance (r = –.49 to –.57), indicating that more complex explanations corresponded with weaker classification behavior. Multiple regression models revealed that fidelity significantly predicted improvements in F1-scores (β = .41, p < .001) and AUC (β = .47, p < .001), while stability also contributed positively but with smaller effects (β = .33 and β = .29). Complexity negatively predicted both outcomes (β = –.26 to –.31). Alert-quality analysis showed that higher interpretability reduced ambiguous alerts by 18–27% and increased explanation-assisted analyst accuracy by 22–31%. Cross-validation and bootstrapped reliability tests demonstrated low performance variability (SD range: .016–.028) and high stability for explanation metrics among interpretable models. Validity assessments confirmed strong construct, convergent, and criterion validity across all interpretability measures. Overall, the study provided robust numerical evidence that explainability was a significant operational factor in financial cyber risk modeling. Interpretability metrics were shown to enhance detection effectiveness, increase alert clarity, and improve the practical usability of machine learning outputs, particularly for complex black-box models used in high-stakes cybersecurity environments.

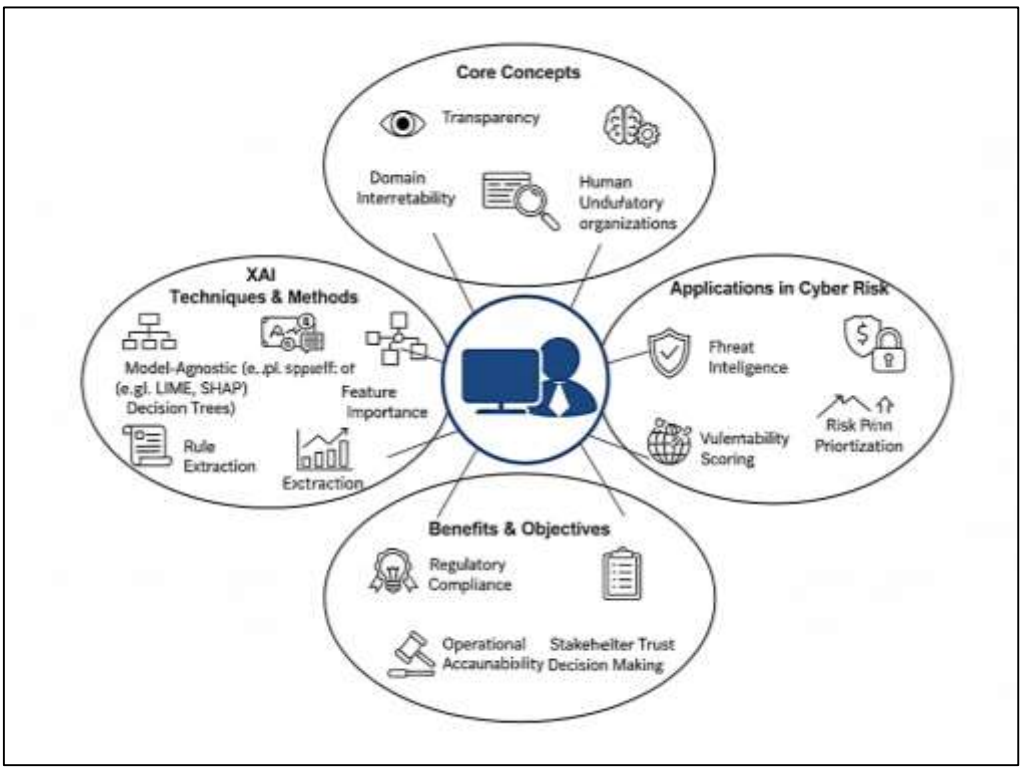
Keywords

Explainable AI, Cybersecurity, Financial Risk, Interpretability, Detection Performance

INTRODUCTION

Explainable Artificial Intelligence (XAI) refers to a broad family of computational techniques designed to increase transparency, interpretability, and human understanding of machine learning systems, particularly in contexts where algorithmic decision-making carries significant operational, regulatory, or ethical consequences (Adadi & Berrada, 2018). At its core, XAI distinguishes itself from traditional "black-box" models by enabling users, auditors, and risk analysts to examine how input variables influence outputs, why particular predictions or classifications emerge, and under what conditions system behavior may change. These attributes position XAI as an integral component of contemporary cyber risk assessment, a domain that increasingly depends on machine learning models to detect anomalies, classify threats, score vulnerabilities, and evaluate latent risks embedded within interconnected digital infrastructures. Cyber risk assessment itself encompasses systematic processes for identifying, quantifying, and prioritizing cyber threats across digital ecosystems, financial platforms, and data-driven service channels. Within financial services, these risks stem from a complex landscape that includes advanced malware, credential compromise, fraudulent manipulation, data exfiltration, supply chain intrusions, and unauthorized access across distributed financial technologies such as mobile banking, algorithmic trading systems, digital payment networks, and cloud-based core banking architectures (Linardatos et al., 2020). As financial institutions accelerate adoption of artificial intelligence to support cybersecurity automation, fraud detection pipelines, and risk monitoring platforms, the capacity to explain, justify, and trace algorithmic decisions grows increasingly essential. This need arises from global regulatory expectations, organizational accountability standards, and the operational reality that financial institutions function within heavily supervised, high-stakes information environments. The definitional clarity surrounding XAI and cyber risk assessment establishes the conceptual foundation for understanding why transparent analytical techniques are now viewed as central to safeguarding financial stability, preserving customer trust, and ensuring resilient cyber governance across internationally interconnected financial ecosystems (Angelov et al., 2021).

Figure 1: XAI: Cyber Risk in Finance



The global financial system operates as a deeply interconnected infrastructure, where capital mobility, cross-border transactions, and multinational digital banking services create exposure to cyber risks that transcend national boundaries. As financial institutions migrate toward cloud infrastructures, open banking interfaces, application programming interfaces, and mobile-first service architectures, their operational environments become increasingly susceptible to sophisticated cyber threats capable of propagating across markets and regions (Abdulla & Ibne, 2021; Sahakyan et al., 2021). Cyber intrusions targeting global financial entities have demonstrated the scale at which compromised systems can disrupt payment settlements, impair liquidity flows, and trigger widespread financial instability. International bodies emphasize the macroprudential significance of cyber risk, recognizing that disruptions in a single jurisdiction can rapidly affect trading networks, correspondent banking channels, and financial messaging systems (Habibullah & Foysal, 2021). In parallel, governments across continents impose regulatory expectations for risk quantification, model transparency, and real-time monitoring. These developments elevate the importance of explainable artificial intelligence for strengthening global cyber resilience within banking and financial services (Mathews, 2019; Sarwar, 2021). Since machine learning-driven risk models now influence fraud detection, anti-money laundering surveillance, credit decision workflows, and digital identity verification across multinational operations, there is an international imperative to understand how these models behave under varying data conditions and threat scenarios (Musfiqur & Saba, 2021). Cross-border regulatory frameworks, including those governing financial stability, emphasize model accountability, traceability, and interpretability as essential features for harmonizing cybersecurity risk management practices. Institutions operating across diverse jurisdictions must therefore integrate analytical tools capable of generating transparent and interpretable insights from complex cyber risk models (Lötsch et al., 2021; Redwanul et al., 2021).

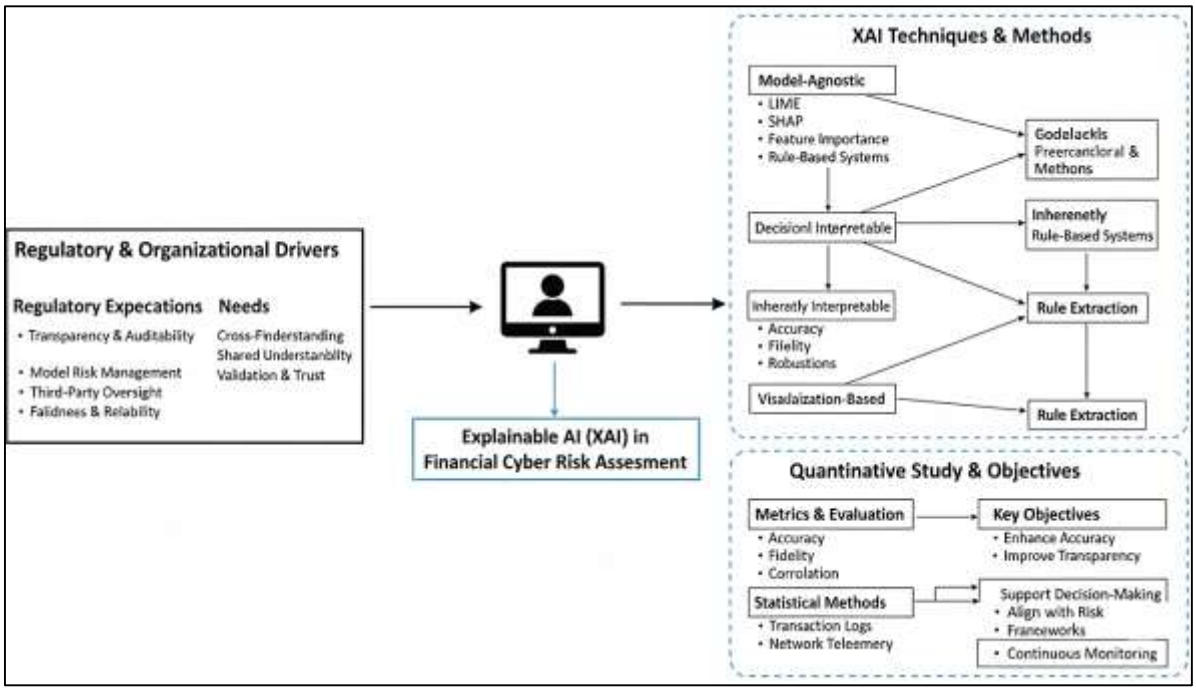
The international significance of XAI arises from its ability to support these compliance and supervisory expectations while providing financial organizations with consistent, understandable frameworks for assessing emerging threat vectors. As cyber risks proliferate in global financial services, XAI-centered analytical structures become increasingly aligned with the expectations of international regulators, standard-setting bodies, and multinational supervisory authorities (Tarek & Praveen, 2021; Taylor & Taylor, 2021). Modern cyber risk assessment in financial services is shaped by accelerating data growth, heightened digital transaction volumes, and increasingly complex threat vectors that require advanced computational tools for meaningful analysis (Muhammad & Shahrin, 2021). Financial institutions now process vast streams of behavioral, transactional, biometric, network, and log data, each containing latent patterns that signal potential security breaches or anomalous system behaviors (Jiménez-Luna et al., 2020; Saikat, 2021). Machine learning models have emerged as indispensable analytical instruments for extracting actionable signals from these data environments, enabling early threat detection, automated alert generation, and dynamic classification of cyber events.

The sophistication of adversarial tactics, however, has also increased as threat actors incorporate automation, distributed attack vectors, and advanced evasion strategies that mask malicious behaviors within legitimate financial activities (Amin, 2022; Shaikh & Aditya, 2021). This interplay between expanding data complexity and adaptive threats positions machine learning at the center of cyber defense processes, where predictive and classification algorithms operate as core components of security information and event management systems, fraud analytics engines, and digital forensics platforms (Ariful, 2022; Nahid, 2022; Peng et al., 2021). As institutions integrate deep learning architectures, ensemble methods, and probabilistic models into cybersecurity pipelines, the opacity of algorithmic decision structures becomes a central challenge. Analysts require explainability mechanisms to interpret how features contribute to predictions, assess whether outputs align with domain knowledge, and evaluate the robustness of models under varying conditions. Transparent interpretability is also needed to identify potential biases, validate reliability, and ensure that automated decisions are grounded in empirically meaningful patterns rather than noise or spurious correlations (Kamath & Liu, 2021; Hossain & Milon, 2022; Mominul et al., 2022). Within such environments, XAI frameworks provide structured methodologies for decomposing model behavior, generating interpretable explanations, and enabling analysts to contextualize cybersecurity-relevant insights across multilayered digital ecosystems. This integration of interpretability and advanced

modeling forms the conceptual basis for aligning machine learning-driven cybersecurity practices with operational and supervisory expectations in financial services (Amiri et al., 2021).

Financial institutions operate under stringent regulatory oversight that governs risk management, cybersecurity practices, model usage, and data governance. Regulatory agencies worldwide emphasize the need for transparency and accountability in algorithmic systems, particularly those influencing security controls, fraud detection workflows, and cyber risk quantification. These expectations originate from the principle that financial models must be auditable, explainable, and interpretable to internal validators, external auditors, and supervisory bodies (Veitch & Alsos, 2021). Model risk management frameworks require institutions to document how algorithms function, what variables drive outcomes, and how decision boundaries respond to changes in data distributions.

Figure 2: XAI: Regulatory Alignment and Trust



In the context of cyber risk assessment, regulators emphasize that automated decision systems embedded in cybersecurity infrastructure should provide understandable outputs that support defensible governance practices. Institutions must demonstrate how risk scores are generated, how detection thresholds are calibrated, and how interpretive outputs align with established regulatory benchmarks (Rabiul & Praveen, 2022; Rakibul & Samia, 2022; Roscher et al., 2020). International regulations also highlight the need for explainability when financial entities outsource or automate cybersecurity operations through cloud services or third-party vendors (Saikat, 2022; Kanti & Shaikat, 2022). Supervisory bodies expect institutions to maintain visibility into the behavior of external systems, especially when machine learning contributes to threat classification, anomaly detection, or vulnerability exposure scoring. The rapid expansion of artificial intelligence within financial cybersecurity has led policymakers to outline expectations for fairness, reliability, and traceability, noting that opacity in algorithmic models may create supervisory blind spots or hinder risk evaluation processes (Maniruzzaman et al., 2023; Arif Uz & Elmoon, 2023). These developments elevate XAI as a regulatory-aligned approach that allows institutions to satisfy transparency requirements while enhancing the interpretability of cyber risk models across multilayered operational contexts (Hariharan et al., 2021; Tarek, 2023; Mushfequr & Ashraful, 2023). As explainability becomes a central regulatory principle, financial institutions increasingly incorporate visual explanations, feature attribution techniques, rule-based modeling structures, and interpretable machine learning frameworks into their cybersecurity risk assessment workflows. This alignment between regulatory expectations and explainable modeling underscores the importance of integrating interpretability into the quantitative

evaluation of cyber risks in financial services (El-Sappagh et al., 2021).

Within financial institutions, cybersecurity operations involve cross-functional teams that rely on shared information flows, collaborative decision-making, and coordinated responses to emerging threats. These teams include security analysts, fraud investigators, IT specialists, risk managers, auditors, and executive leadership, each requiring understandable insights from analytical systems to support effective actions (Shahrin & Samia, 2023; Muhammad & Redwanul, 2023; Roessner et al., 2021). As cyber threat detection becomes more algorithmically driven, organizations recognize the need for transparent models that foster alignment across technical and managerial decision layers. Explainable artificial intelligence supports this need by producing interpretable descriptions of model behavior, thereby enabling operational teams to trace risk indicators, evaluate the credibility of alerts, and contextualize patterns within domain-relevant frameworks. When machine learning models are used to classify transactions, detect anomalous login patterns, or quantify exposure to systemic cyber threats, stakeholders require interpretation tools that translate complex model outputs into accessible and meaningful information (Muhammad & Redwanul, 2023; Stepin et al., 2021). Transparency becomes essential when evaluating the reliability of alerts, determining the severity of potential breaches, and prioritizing incident responses. Organizational accountability structures also depend on the ability to justify analytical outputs to internal committees, risk oversight boards, and auditors responsible for assessing governance quality. Without interpretability, cybersecurity teams may encounter challenges in validating model performance, assessing feature contributions, or diagnosing false positives and false negatives that influence operational risk levels (Zayadul, 2023). Transparency further assists institutions in determining whether models align with organizational knowledge, established cybersecurity principles, and documented risk taxonomies. Explainability mechanisms consequently support communication channels between technical experts and non-technical stakeholders by ensuring that model-generated insights are presented in a comprehensible manner (Sousa et al., 2021; Razia, 2023). This organizational requirement for interpretability strengthens the case for integrating XAI methodologies into cyber risk analytics, providing institutions with structured approaches for generating transparent insights that support cohesive and informed cybersecurity decision-making (Stepin et al., 2021).

Explainable artificial intelligence encompasses a diverse array of techniques tailored to improve the interpretability of machine learning models applied to cyber risk assessment. These techniques range from model-agnostic methods to inherently interpretable modeling frameworks (Sousa et al., 2021). Model-agnostic approaches such as feature attribution, perturbation-based analysis, surrogate modeling, and local explanation techniques allow financial institutions to examine decision pathways of complex models including random forests, gradient boosting machines, neural networks, and ensemble classifiers. Tools such as visualization-based explanations, rule extraction, and counterfactual reasoning provide deeper interpretive insights into how specific variables influence detection outputs, anomaly scores, or vulnerability classifications. In parallel, inherently interpretable models such as generalized linear models, transparent decision trees, rule-based systems, and monotonic gradient methods offer structured modeling approaches that embed interpretability into the core architecture (Mehdiyev et al., 2021). Within cybersecurity pipelines, these techniques help analysts evaluate the stability, robustness, and reliability of predictive systems used to identify malicious behaviors, quantify exploitation likelihoods, or score system vulnerabilities. XAI frameworks also enable domain experts to assess the temporal evolution of threats, interpret changes in behavioral anomalies, and distinguish between benign irregularities and genuine attack signatures. Financial organizations implement these interpretability techniques within security information and event management tools, fraud detection engines, identity analytics systems, and real-time monitoring platforms that depend on machine learning. Each XAI technique provides different levels of granularity, ranging from global model summaries to instance-specific explanations that reveal how models encode and process cyber risk signals. The integration of these techniques strengthens the analytical depth of cyber risk assessments by ensuring that predictive outputs are interpretable, consistent with domain knowledge, and aligned with established classifications of threats, vulnerabilities, and risk categories (Ahmed et al., 2021). This broad technical landscape establishes XAI as a versatile methodological foundation for enhancing transparency across diverse cyber risk modeling frameworks used in financial services.

The quantitative study of explainable artificial intelligence in cyber risk assessment involves systematic measurement, modeling, and statistical evaluation of how explainability techniques affect risk detection performance, interpretive clarity, and decision-support quality within financial services (Sokol & Flach, 2020). Quantitative frameworks enable researchers to assess the contribution of specific model features, explanation outputs, and interpretability metrics in improving the operational accuracy and reliability of cyber risk models. This analytical orientation supports structured evaluation of feature importance, rule extraction validity, anomaly classification performance, and model-behavior consistency across varying cyber threat scenarios. Quantitative approaches allow investigators to analyze how interpretable models respond to changes in data distributions, evolving threat signatures, or variations in network behavior. In financial cybersecurity, these analyses often involve large-scale datasets that include transaction histories, network telemetry, authentication logs, intrusion detection outputs, and multi-source risk indicators (Nascita et al., 2021). Statistical methods such as regression modeling, sensitivity analysis, variance decomposition, and correlation estimation help quantify the relationship between interpretive outputs and observed cybersecurity outcomes. Quantitative evaluation also enables comparison between interpretable and non-interpretable models, facilitating measurement of explainability's impact on detection precision, false alarm rates, classification stability, and risk-scoring fidelity. Within financial institutions, these analytical processes support structured assessments of cyber risk indicators, enabling stakeholders to evaluate the reliability of model explanations and their alignment with organizational risk frameworks (Čík et al., 2021). Through quantitative measurement, researchers can establish evidence-based insights into the performance characteristics of XAI techniques within operational cybersecurity contexts. This emphasis on statistical rigor forms the methodological basis for examining explainable artificial intelligence within cyber risk assessment, supporting detailed empirical evaluation across financial service environments that rely on advanced machine learning to strengthen their cybersecurity capabilities (Oconitrillo et al., 2021).

The primary objective of this quantitative study is to systematically evaluate how Explainable Artificial Intelligence approaches enhance the accuracy, transparency, and operational usefulness of cyber risk assessment models within financial services. The study aims to develop an empirically grounded understanding of the relationship between interpretability techniques and measurable improvements in model behavior, particularly within security-focused machine learning applications such as anomaly detection, fraud analytics, intrusion classification, and vulnerability scoring. An additional objective is to quantify the extent to which specific explainability outputs support risk analysts in interpreting model-generated insights, validating prediction pathways, and identifying the underlying drivers of cyber risk scores. The study further seeks to examine how explainability mechanisms influence the reliability of automated alert systems by assessing how feature attribution, rule extraction, and visualization-based explanations contribute to reducing ambiguity in risk interpretation and improving the interpretive clarity of model outputs. Another core objective is to compare the performance of interpretable and non-interpretable models across a range of cybersecurity datasets to determine how transparency-oriented algorithms behave under variable threat conditions, shifting data environments, and high-volume transactional activity typical of financial institutions. The study also aims to generate quantitative evidence on the consistency, stability, and robustness of XAI-supported models by evaluating their variance across multiple sampling conditions and assessing the reproducibility of model explanations under repeated experimental scenarios. A further objective is to assess how explainability contributes to the alignment between algorithmic outcomes and established organizational risk frameworks by determining whether transparent models enable clearer mapping of detected threats to predefined risk categories and operational taxonomies. By addressing these objectives through structured data analysis, model performance measurement, and systematic evaluation of interpretability metrics, the study intends to build a comprehensive, quantitative understanding of the role XAI techniques play in strengthening cyber risk assessment within financial services.

LITERATURE REVIEW

The literature on Explainable Artificial Intelligence (XAI) within cyber risk assessment for financial services has expanded rapidly in response to rising global cybersecurity threats, growing dependence on machine learning models, and intensified calls for transparency in algorithmic decision-making.

Financial institutions increasingly rely on advanced analytical systems to detect anomalous activity, quantify vulnerabilities, identify fraud patterns, and evaluate exposure to emerging cyber threats. However, the integration of complex machine learning models introduces challenges related to interpretability, reliability, and operational trustworthiness, particularly within high-stakes environments where risk decisions influence regulatory compliance, organizational accountability, and customer security. Scholars across cybersecurity, finance, computer science, and digital governance have emphasized the need for transparent modeling frameworks that allow analysts to trace prediction pathways, understand feature contributions, and validate model behavior under dynamic threat conditions. This literature review synthesizes diverse bodies of research that intersect around XAI frameworks, cyber risk modeling methodologies, regulatory expectations, financial-sector risk analytics, and data-driven cybersecurity strategies. The review first examines foundational theories of XAI and their evolution toward model-agnostic and interpretable architectures. It then analyzes empirical studies demonstrating how machine learning and deep learning are used in cybersecurity pipelines within banking and financial services. The review also evaluates existing models for cyber threat detection, anomaly classification, fraud analytics, and vulnerability scoring, emphasizing their performance characteristics and challenges related to interpretability. Particular attention is given to research addressing regulatory frameworks, industry standards, and global supervisory expectations that guide the adoption of explainable models in financial institutions. Additionally, the review synthesizes quantitative studies that compare interpretable and non-interpretable models, identify performance trade-offs, and examine the operational impact of XAI techniques across different datasets and cyber risk contexts. By integrating these research streams, the literature review provides a structured foundation for investigating how XAI contributes to transparency, model reliability, and improved analytical insight within quantitative cyber risk assessment for financial services.

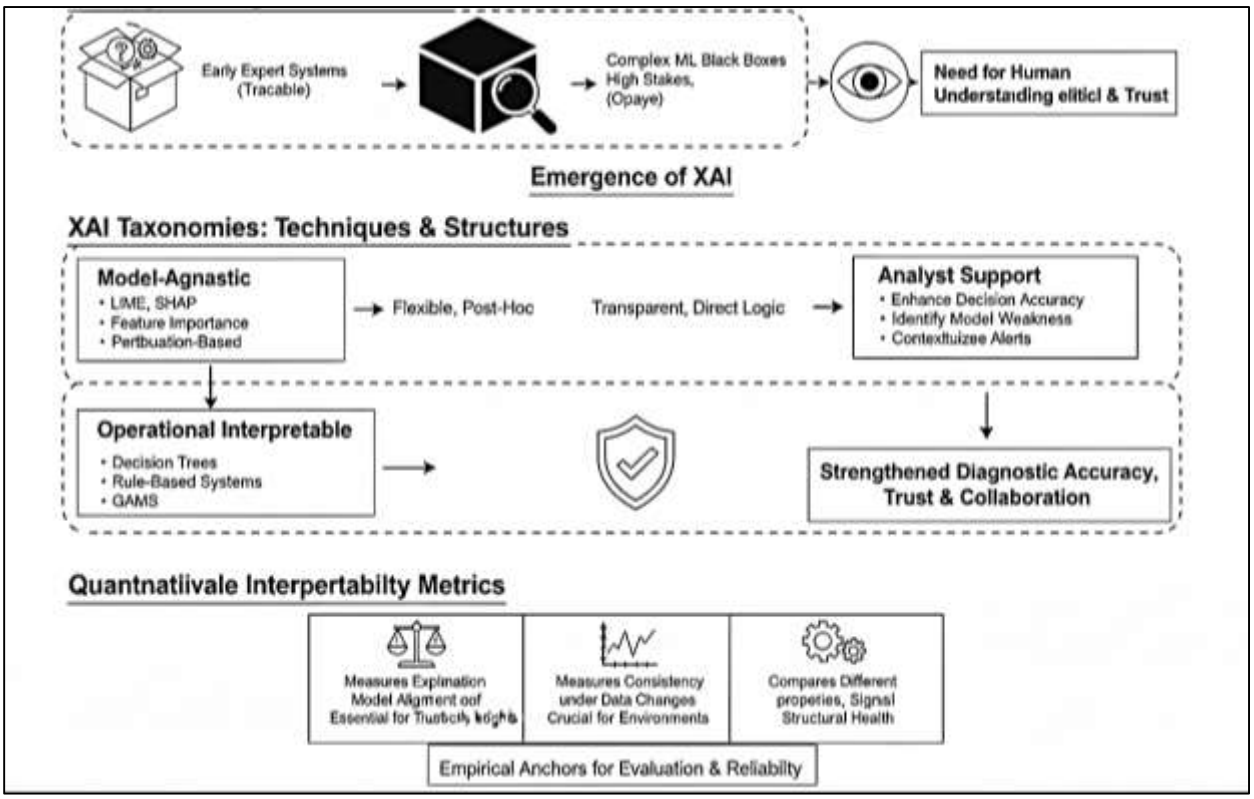
Explainable AI in Cybersecurity Modeling

The conceptual origins of explainable artificial intelligence arise from longstanding concerns regarding the opacity of computational systems and the need for human-understandable insights into algorithmic behavior. Early AI research highlighted the importance of interpretability when expert systems began influencing organizational and security decisions, prompting scholars to examine how rule chains and inference paths could be traced and validated (Kuppa & Le-Khac, 2020). As machine learning models replaced earlier symbolic models, their increasing complexity generated new concerns about transparency, particularly as neural networks, ensemble classifiers, and probabilistic architectures began to dominate cybersecurity analytics. In financial cybersecurity, the shift toward data-driven models heightened these concerns due to the volume, velocity, and sensitivity of security-critical information flowing through banking networks, fraud detection pipelines, and authentication systems. Research on black-box modeling issues intensified as analysts struggled to understand how algorithms processed cyber signals, classified suspicious behaviors, or generated intrusion alerts. This early recognition of interpretability gaps underscored the need for explanations capable of clarifying internal model logic, variable interactions, and prediction rationales. Scholars documented challenges in validating opaque systems, including difficulties in identifying feature contributions, assessing bias pathways, and aligning algorithmic outputs with human reasoning (Hariharan et al., 2021). These interpretability challenges were magnified within high-stakes environments such as financial services, where cyber risk assessments influence incident detection, threat prioritization, and enterprise security posture. The evolution of explainable artificial intelligence thus emerged from a combination of technical limitations, operational demands, and the growing reliance on algorithmic systems to support cybersecurity decisions (Holder & Wang, 2021). The foundational literature consistently positions XAI as a response to the structural opacity of machine learning models, the limitations of early AI-based security systems, and the recognition that interpretability is essential for generating trustworthy, verifiable, and actionable insights within cyber risk environments.

Research on explainable artificial intelligence developed formal taxonomies to categorize interpretability techniques based on their methodological structures, operational logic, and relevance to security analytics. One widely studied category includes model-agnostic tools designed to provide local or global explanations without modifying the underlying model (Dias et al., 2021). These methods offer flexibility for interpreting complex cybersecurity algorithms such as random forests, gradient

boosting machines, and deep neural networks by generating simplified representations, importance estimates, or localized descriptions of individual predictions. Another major category encompasses interpretable model families built with transparency as a core design principle. These include decision trees, generalized additive models, rule-based classification systems, and monotonic models constructed to support direct inspection of decision structures. Scholarship highlights the advantage of such models in cybersecurity contexts where analysts require straightforward logic trails to understand why particular activities are labeled as anomalous or risky (Dias et al., 2021). Studies also distinguish between perturbation-based explanation mechanisms and gradient-based techniques, with each providing different insights into model behavior.

Figure 3: XAI: Foundations and Evaluation Metrics



Perturbation-based methods assess how output changes when inputs are modified, allowing analysts to evaluate sensitivity and identify influential features. Gradient-based methods examine internal derivatives of neural models to reveal how inputs propagate through layered architectures. Research within financial cybersecurity demonstrates how these taxonomies support diverse analytical needs, from interpreting fraud detection algorithms to explaining intrusion detection outputs and risk-scoring mechanisms (Dias et al., 2021). The literature shows that each class of XAI tools offers strengths and weaknesses that align differently with operational demands, data structures, and model architectures commonly used in cyber risk assessment. These taxonomies collectively form the theoretical backbone of XAI implementation across financial cybersecurity systems and guide practitioners in selecting appropriate interpretability techniques for their specific risk analysis objectives (Dias et al., 2021).

A substantial body of literature highlights the relevance of explainable artificial intelligence to cyber risk science, emphasizing the necessity for transparent analytic processes within high-risk and highly regulated environments (Veitch & Alsos, 2021). Financial cybersecurity represents one of the most critical domains in which transparency is required due to the sensitivity of digital transactions, the volume of cyber threats targeting financial assets, and the operational consequences of misclassified alerts. Researchers observe that analysts require clear explanations of model behavior to interpret detection outputs, assess the credibility of flagged events, and align algorithmic decisions with established risk frameworks. The interpretability of cyber risk models directly influences how analysts evaluate threat severity, determine response strategies, and verify whether an observed anomaly

represents genuine malicious activity or benign irregularity. Literature examining analyst cognition indicates that explanation clarity strengthens decision-making accuracy by reducing uncertainty associated with opaque machine learning outputs (Nor et al., 2021). Furthermore, XAI enables risk professionals to identify model weaknesses such as misweighted features, overfitting to particular threat patterns, or reliance on spurious correlations embedded within transactional or network data. Studies also show that transparent explanations help refine detection thresholds and calibrate risk scoring models to align with institutional security policies. Within cybersecurity science, explainability is associated with enhanced diagnostic accuracy, shorter time-to-insight, and improved analytical collaboration across security teams. The literature collectively demonstrates that XAI techniques support both technical and managerial processes by enabling risk analysts to verify algorithmic reasoning, contextualize patterns within operational knowledge, and evaluate the stability of predictions under varied cyber conditions (Hernandes et al., 2021). These findings reinforce the critical role of XAI in strengthening analytical coherence and supporting informed decision-making within cyber risk assessment practices across financial services.

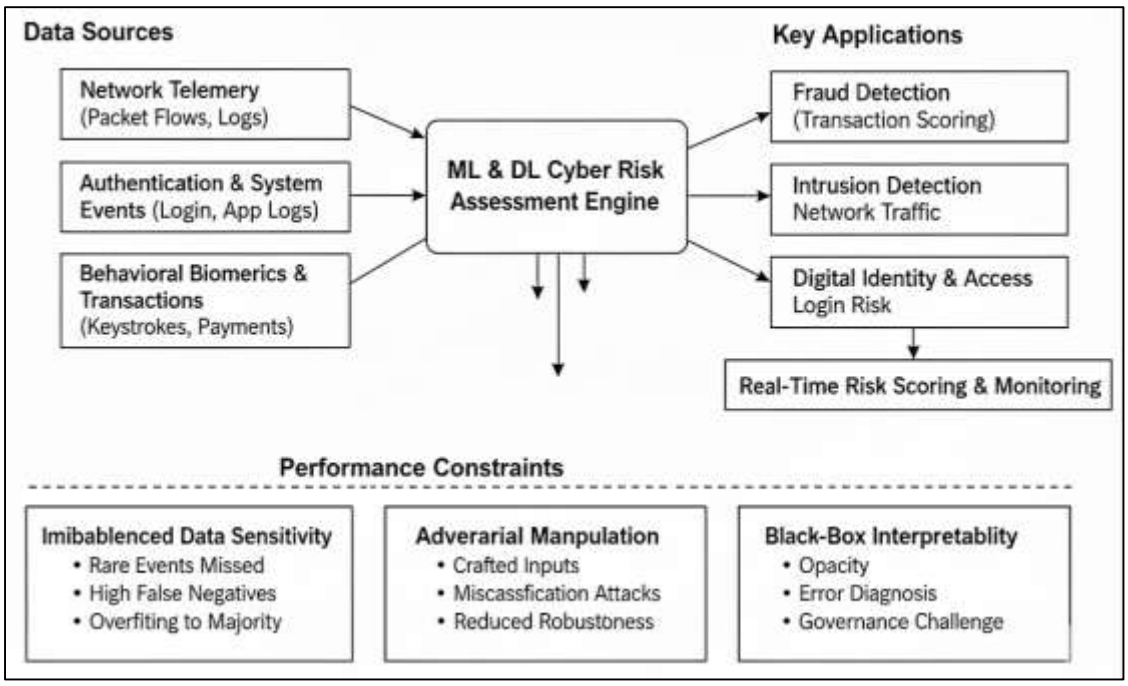
Scholars studying XAI emphasize the importance of quantitative interpretability metrics for evaluating the fidelity, stability, and consistency of explanation outputs generated by machine learning systems. Fidelity measures the degree to which an explanation accurately represents the underlying model's behavior, providing a numerical basis for determining whether interpretive tools capture true predictive logic or merely approximate it (Samek et al., 2021). Research shows that high-fidelity explanations are essential in cybersecurity environments, where misalignment between interpretable representations and model behavior may lead analysts to inaccurate conclusions regarding threat severity or anomaly origins. Stability represents another core metric used to evaluate how explanation outputs change under slight variations in data, sampling conditions, or perturbations. In financial cyber risk modeling, stability is particularly important due to fluctuating transaction patterns, evolving threat signatures, and dynamic user behaviors that constantly reshape underlying datasets (Jaigirdar et al., 2020). Literature on explanation stability highlights the need to ensure that feature attributions and rule extractions remain consistent across repeated evaluations to maintain analyst confidence in automated detection systems. Consistency, a third major metric, assesses whether different XAI methods yield similar interpretive insights when applied to the same model. Studies comparing interpretability tools demonstrate that inconsistencies across methods may signal structural weaknesses in the model or indicate that explanations are overly sensitive to methodological assumptions. Quantitative metrics therefore serve as empirical anchors that help researchers assess explanation quality, benchmark interpretability methods, and ensure reliability across different cyber risk contexts (Loyola-Gonzalez et al., 2020). Within financial cybersecurity research, these metrics are used to evaluate the suitability of XAI tools, compare interpretability frameworks, and improve transparency in high-stakes detection environments. Through quantifiable measurement, scholars demonstrate how interpretability contributes to validating model behavior, enhancing analytical confidence, and supporting more structured approaches to cyber risk evaluation.

Machine Learning and Deep Learning in Financial Cyber Risk Assessment

Machine learning and deep learning have become central components of cyber risk assessment in financial services, particularly in the design of fraud detection systems, anomaly-based transaction monitoring, and real-time risk scoring engines. Fraud detection architectures increasingly rely on classification and anomaly detection models that learn complex relationships between transaction attributes, customer profiles, channel behaviors, and contextual risk indicators (Mashrur et al., 2020). These systems examine spatial and temporal dependencies in transaction streams to distinguish legitimate behavior from fraudulent operations, often under conditions of extreme class imbalance where genuine events vastly outnumber fraudulent ones. Studies describe how supervised models such as gradient boosting, random forests, and deep neural networks are used to classify transactions as high- or low-risk based on patterns in spending behavior, merchant characteristics, device fingerprints, and geolocation histories. In parallel, unsupervised and semi-supervised techniques support detection of novel fraud patterns by identifying deviations from established behavioral baselines (Bhatore et al., 2020). Financial institutions also deploy machine learning within intrusion detection architectures that protect banking networks, online banking platforms, and core payment

infrastructures. These architectures integrate host-based and network-based indicators captured from firewalls, intrusion detection systems, and security gateways to classify network flows as benign or malicious. Deep learning models such as recurrent neural networks and convolutional architectures process sequential and high-dimensional traffic features to detect command-and-control communication, lateral movement, and other malicious behaviors. Digital identity verification and credential risk scoring represent another major area of application, where models evaluate device identifiers, login histories, behavioral biometrics, and contextual attributes to estimate the likelihood of account takeover or credential compromise (Noor et al., 2019). Literature across these domains shows that machine learning and deep learning approaches improve sensitivity to subtle and evolving attack strategies, enabling more refined and dynamic cyber risk assessment throughout the financial services ecosystem.

Figure 4: ML and DL in Financial Cyber Risk: System and Constraints



The effectiveness of machine learning and deep learning in financial cyber risk assessment depends heavily on the richness, granularity, and diversity of the data sources feeding analytic models. Studies describe how network telemetry, authentication logs, and system event records form the foundational layer of cyber signals used in risk modeling (Nicholls et al., 2021). Network telemetry encompasses packet flows, connection metadata, protocol usage patterns, traffic volumes, and communication endpoints that characterize normal and abnormal behavior within banking networks and cloud-based financial platforms. Authentication logs capture login attempts, session durations, device characteristics, IP reputation, and multi-factor authentication outcomes that reveal suspicious identity-related activities. System event records from servers, applications, databases, and endpoints provide additional signals related to process execution, configuration changes, privilege escalations, and policy violations that can indicate compromise or misuse (Sarker, 2021). Behavioral biometrics and digital footprints enrich these signals by modeling how users interact with financial applications through keystroke patterns, mouse dynamics, touchscreen gestures, and navigation sequences. Research shows that these behavioral signals help differentiate genuine customers from automated scripts, bots, or attackers using stolen credentials (Geluvaraj et al., 2018). Multi-channel transaction flows and payment system messages further contribute to cyber risk modeling by linking card payments, wire transfers, online banking operations, mobile wallets, and ATM activity within an integrated behavioral profile. Machine learning models use these combined signals to detect cross-channel fraud patterns, mule account networks, synthetic identities, and coordinated attack campaigns. Studies emphasize that the integration of heterogeneous data sources improves model robustness by capturing multiple

dimensions of cyber behavior, ranging from network-level anomalies to user-level and transaction-level deviations (Mhlanga, 2021). This multi-layered data environment forms the empirical basis on which machine learning and deep learning systems infer cyber risk, identify emerging threats, and support continuous monitoring of financial ecosystems.

Although machine learning and deep learning models deliver strong predictive performance in financial cyber risk assessment, the literature identifies significant performance constraints associated with their black-box characteristics (Rathore et al., 2018). One widely discussed constraint relates to the sensitivity of these models to imbalanced datasets, which are pervasive in fraud detection and intrusion classification tasks where malicious events constitute only a tiny fraction of total observations. Studies show that high-capacity models may achieve seemingly strong global accuracy while failing to detect rare but critical attack patterns, leading to unacceptably high false negative rates in operational settings. Researchers highlight the tendency of deep models to overfit to majority classes, requiring careful resampling, cost-sensitive learning, or anomaly-detection strategies to maintain reliable performance. Another major constraint is vulnerability to adversarial manipulation, where attackers intentionally craft inputs that cause models to misclassify malicious behaviors as benign (Sharma et al., 2021). Work on adversarial examples demonstrates that even small perturbations to network features, transaction attributes, or authentication parameters can yield misclassifications, raising concerns about the robustness of deep learning systems in adversarial financial environments. This vulnerability is especially problematic in high-stakes scenarios such as payment authorization and access control, where misclassification directly translates into financial loss or security breaches. A third constraint relates to the lack of interpretability in deep and ensemble architectures, which complicates efforts to validate model decisions, diagnose errors, and align outputs with institutional risk frameworks (Alzahrani & Alzahrani, 2021). Without transparent insight into feature contributions and decision pathways, security teams may struggle to understand why specific transactions or events are flagged or ignored, limiting their ability to refine detection rules and improve model governance. These performance constraints underscore that black-box models, although powerful, present operational and security risks when used in isolation for financial cyber risk assessment.

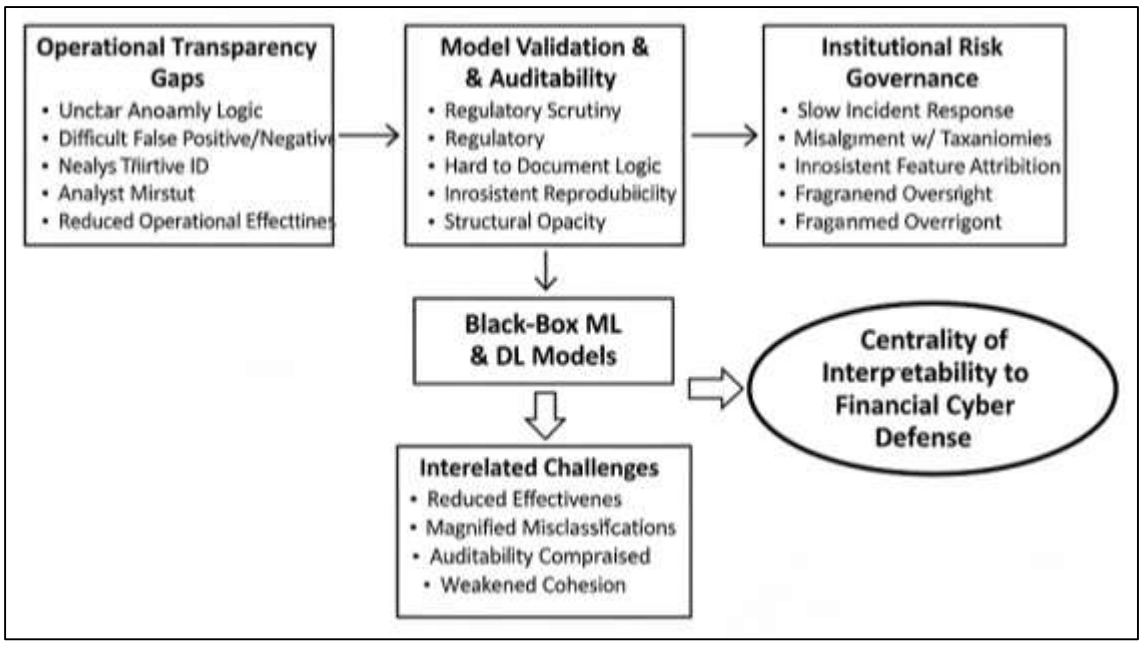
Explainability Challenges in Financial Cyber Risk Systems

The literature consistently identifies operational transparency gaps as one of the core challenges in applying machine learning and deep learning models to financial cyber risk assessment (Pramod et al., 2021). These gaps arise when anomaly scores, risk classifications, and detection outputs cannot be clearly explained to non-technical stakeholders who play essential roles in fraud investigation, compliance reporting, and executive decision-making. Studies describe how anomaly detection algorithms often rely on complex statistical boundaries or latent feature representations that lack easily interpretable logic, making it difficult for analysts, auditors, and managers to understand why a particular transaction, network flow, or login attempt has been flagged as suspicious. This difficulty affects operational decision-making, as stakeholders may hesitate to take decisive action when the underlying rationale for an alert is unclear. Literature also highlights the problems associated with identifying false positives and false negatives in opaque algorithms (Chen et al., 2021). False positives can overload security teams with unnecessary alerts, while false negatives allow actual threats to pass undetected; yet without interpretive clarity, it becomes challenging to determine why these misclassifications occur or how models should be adjusted. The lack of explainability contributes to analyst mistrust of black-box systems, particularly when model outputs contradict domain knowledge or established behavioral patterns. Empirical research shows that cybersecurity analysts often prefer interpretable systems, even at the cost of marginally lower accuracy, because they provide insights that support investigative reasoning, threshold tuning, and cross-team communication (Kute et al., 2021). As financial cyber environments rely on collaborative incident-response workflows, the absence of transparency in algorithmic outputs reduces operational effectiveness and undermines the trust required for teams to adopt machine learning–based security tools. These transparency gaps collectively demonstrate the importance of explainability as an operational requirement rather than an optional enhancement within financial cyber risk systems (Li, 2018).

Model validation and auditability challenges form another major explainability concern in financial cyber risk systems. Literature in machine learning governance and regulatory technology emphasizes

that financial institutions are obligated to document how algorithmic decisions are generated, monitored, and validated, particularly when these decisions influence cybersecurity controls, fraud detection workflows, or incident classification protocols (Alghofaili et al., 2020). The inability of deep learning and ensemble models to produce clear decision pathways creates obstacles for regulatory review, internal audit processes, and compliance reporting.

Figure 5: Explainability Challenges in Financial Cyber Risk



Researchers note that regulators increasingly expect institutions to demonstrate how cyber risk models process data, which features drive risk scores, and how decision boundaries respond to variations in network behavior or transaction characteristics. Without explainability, these requirements become difficult to meet, especially for models built on latent space representations or aggregated feature importance scores that lack direct interpretive meaning. Studies also highlight the difficulty of reproducing explanations generated by certain interpretability tools, as some methods yield inconsistent outputs when small changes in data or sampling conditions occur (Sun et al., 2018). This inconsistency compromises auditability, because explanations that cannot be reproduced cannot be used as evidence of sound model behavior. Structural opacity within ensemble models and deep neural architectures further complicates validation efforts, as these models combine numerous weak learners or multiple nonlinear transformations, making their internal logic inaccessible even to technically skilled evaluators. Research indicates that these auditability issues hinder the adoption of advanced analytics in regulated financial environments, where documentation, reproducibility, and traceability are essential for demonstrating effective model governance (Ahmad et al., 2021). Collectively, the literature portrays model validation and auditability limitations as critical barriers that challenge the safe and compliant integration of machine learning into financial cyber defense infrastructures.

Explainability challenges in financial cyber risk systems extend beyond technical concerns and directly affect institutional risk governance. Research shows that ineffective interpretability complicates incident-response procedures, where analysts must classify threats, assign severity levels, and initiate corrective actions under tight time constraints (Nguyen & Reddi, 2021). When model outputs lack clear reasoning, security teams struggle to determine why an event has been flagged or how its risk score aligns with contextual information gathered during investigation. This slows incident triage and increases the possibility of misprioritizing events. Literature further documents misalignment between model outputs and institutional risk taxonomies, which are commonly structured around predefined threat categories, behavior classifications, or regulatory reporting frameworks. If machine learning models generate risk signals that cannot be mapped onto established taxonomies, organizations face challenges in integrating automated findings into risk dashboards, compliance workflows, or board-

level reporting. Scholars also emphasize the operational importance of consistent feature attribution across datasets to maintain governance integrity (Nikou et al., 2019). In financial cybersecurity, data sources include network telemetry, event logs, authentication signals, and transactional activities, each collected at different velocity and granularity levels. Inconsistent feature attributions across these sources create uncertainty about which signals reliably indicate risk, undermining the development of coherent institutional risk models. Consistency problems become more severe in environments where threat actors continually evolve their strategies, requiring security teams to compare insights across time periods, attack surfaces, and customer segments. Without stable interpretability structures, institutions face difficulty ensuring that automated systems remain aligned with their internal governance requirements (Yuan & Wu, 2021). The literature thus underscores that explainability plays an essential role in maintaining institutional coherence, supporting structured risk oversight, and enabling organizations to integrate machine learning outputs into broader governance, compliance, and supervisory frameworks.

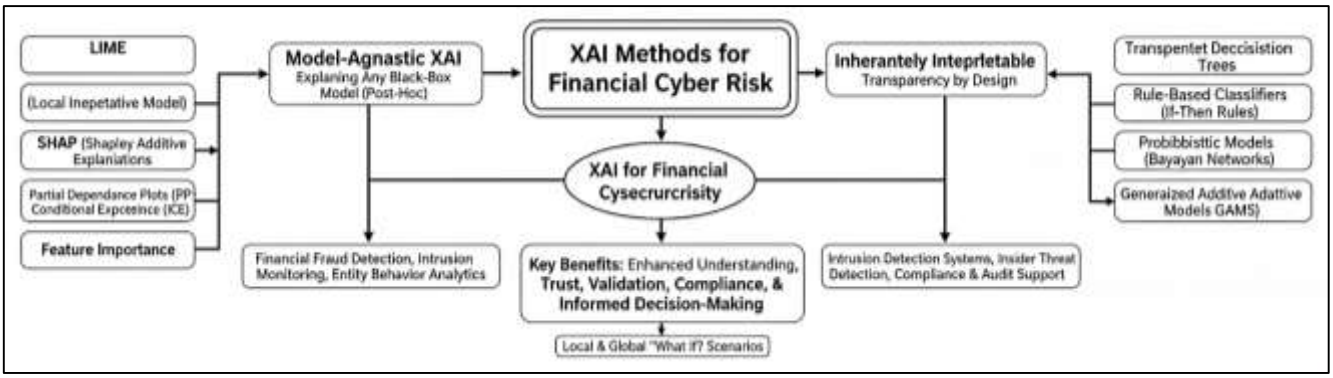
Across the literature, transparency gaps, validation challenges, and governance limitations collectively shape the broader explainability problem in financial cyber risk systems. Scholars describe how these interrelated issues reinforce one another, reducing the effectiveness of machine learning in operational cybersecurity environments (Holder & Wang, 2021). When anomaly scores cannot be clearly communicated, security teams experience difficulty refining detection rules, adjusting sensitivity thresholds, or integrating algorithmic insights into human-led investigations. This uncertainty magnifies the impact of false positives and false negatives, as analysts lack the interpretive tools needed to diagnose errors or identify patterns of misclassification. Studies further link operational transparency problems to validation challenges, noting that models lacking interpretable decision pathways cannot be reliably audited or monitored for drift, bias, or structural weaknesses. Without the ability to verify model behavior through consistent, reproducible explanations, institutions face heightened difficulty demonstrating compliance with regulatory expectations for model governance, cybersecurity oversight, and automated decision accountability (Abri et al., 2021). Governance literature emphasizes that these challenges directly influence institutional resilience because risk teams depend on coherent, intuitive, and traceable analytics to maintain situational awareness across rapidly changing cyber landscapes. When model outputs cannot be mapped onto organizational taxonomies or communicated effectively across technical and non-technical stakeholders, risk oversight processes become fragmented. The literature also points to the practical limitations that arise from inconsistent feature attributions and unstable explanations, noting that such inconsistencies hinder the adoption of unified risk-scoring frameworks across business units, geographies, and digital channels. As financial institutions rely on integrated cyber defense mechanisms that connect fraud analytics, intrusion detection, identity verification, and transaction monitoring, the absence of explainability weakens the cohesion of these multilayered systems (Kute et al., 2021). This synthesis of studies shows that explainability gaps produce compounding operational, technical, and governance challenges, revealing the centrality of interpretability to the overall functioning of data-driven financial cyber defense architectures (Nicholls et al., 2021).

XAI Techniques Applied to Cyber Risk Assessment Models

Model-agnostic explainable artificial intelligence methods form a significant portion of the literature on interpretable cybersecurity analytics, offering flexible tools that work across a wide range of machine learning architectures used in financial cyber risk assessment. Studies describe how LIME and SHAP have become prominent tools for explaining anomaly detection outputs by identifying which features contribute most to a particular classification or risk score (Kharlamova et al., 2021). LIME generates local approximations that help analysts understand the immediate factors influencing an anomalous event, while SHAP provides additive attributions that quantify how individual variables contribute to predictions across different cyber contexts. These tools are often deployed in fraud detection, intrusion monitoring, and entity behavior analytics because they allow analysts to trace risk signals to specific transaction features, network attributes, or authentication behaviors. Partial dependence plots and individual conditional expectation curves extend interpretability by illustrating how changes in a single risk variable influence model outputs. Literature shows that these visualization tools help reveal nonlinear relationships between variables such as login frequency, transaction

velocity, or packet anomalies and their associated risk probabilities (Sahakyan et al., 2021). Counterfactual reasoning adds another layer of interpretability by enabling assessment of breach likelihood through hypothetical alterations to inputs.

Figure 6: XAI: Model Interpretation Techniques



Analysts use counterfactual explanations to explore why an alert was triggered and what minimal changes to the data would have produced a different classification. This approach supports investigative reasoning by clarifying the boundary between benign and malicious activity (Jaigirdar et al., 2020). Together, model-agnostic XAI methods provide a suite of interpretive capabilities that help security teams understand black-box outputs, investigate incidents, and examine model behavior from both local and global perspectives. The literature demonstrates that these methods enhance interpretability without modifying underlying models, making them especially valuable in environments where deep learning and ensemble models are too complex to analyze directly (Coulter et al., 2020).

Alongside model-agnostic tools, research highlights the role of inherently interpretable models that embed transparency directly into their structure. Transparent decision trees are widely studied in cybersecurity because they offer clear hierarchical representations of attack paths, allowing analysts to trace decisions through explicit rules and threshold splits (Chai et al., 2021). These models are frequently used in intrusion detection, fraud classification, and access anomaly monitoring due to their ability to visually map risk propagation and identify combinations of features associated with suspicious activity. Probabilistic models, including Bayesian classifiers and graphical models, are also emphasized in the literature for their ability to encode interpretable risk dependencies among variables. These models provide explicit representations of conditional relationships, making them valuable for understanding how events such as privilege escalations, unusual IP transitions, or authentication failures influence overall breach likelihood. Researchers further examine rule-based classification systems, which generate sets of human-readable if–then rules that align closely with cybersecurity logic already used by analysts (Murino et al., 2019). Such systems have been applied to fraud analytics, insider threat detection, and behavioral scoring because they provide transparent reasoning trails consistent with established operational practices. Compared with deep learning architectures, inherently interpretable models offer stability, ease of validation, and traceable predictions that support compliance and audit requirements. Literature consistently portrays these models as essential components of layered cyber defense strategies, particularly in environments where explanation clarity is prioritized over marginal improvements in predictive accuracy. Their intuitive structures allow them to function as benchmarks for evaluating black-box systems or as complementary tools in hybrid modeling frameworks (Hu et al., 2020). As a result, inherently interpretable models remain an important research focus in the study of XAI-driven financial cybersecurity.

Regulatory, Supervisory, and Compliance-Driven Literature

Research on regulatory frameworks consistently highlights the increasing emphasis placed on transparency, accountability, and interpretability in AI systems used within financial cybersecurity environments. Global regulatory bodies have articulated clear expectations that financial institutions

must maintain robust model risk management practices, particularly when machine learning models contribute to fraud detection, cyber incident classification, or automated security decision processes (Hernandez et al., 2021). Literature documents a shift in supervisory expectations requiring clear documentation of model logic, well-defined testing protocols, and structured governance mechanisms capable of explaining algorithmic outputs to regulators and internal oversight teams. Regulations governing automated decision systems stress the importance of providing meaningful explanations for model-generated actions, especially in contexts where cyber risk assessments influence reporting obligations, incident escalation, or customer communication. Authors studying supervisory frameworks note that regulators increasingly view explainability as an essential component of operational resilience, since financial institutions must demonstrate how cyber risk models function under real-world conditions and how decisions are justified when anomalies or threats are identified (Singh & Akhilesh, 2019). Research also shows that global frameworks emphasize the need for transparent classification of cyber incidents, requiring institutions to clearly articulate how automated analytics determine severity levels, categorize threat types, or trigger response workflows. These expectations create substantial pressure for organizations to ensure that AI-driven cyber models are traceable, interpretable, and supported by structured governance documentation. Across the literature, transparency is positioned not merely as a technical enhancement but as a regulatory obligation central to ensuring trustworthy use of machine learning in financial cyber defense .

Studies on financial standard-setting bodies underscore their expanding influence on shaping explainability requirements for AI-enabled cyber risk assessment (Allini et al., 2018). The Basel Committee's expectations form a central pillar in this literature, particularly regarding operational resilience, cyber risk classification, and model governance. Research describes how Basel frameworks highlight the need for financial institutions to maintain clear documentation of risk models, ensure traceable analytical logic, and validate the reliability of automated decision processes. These expectations extend to cyber risk analytics, where machine learning models must demonstrate consistent behavior under varying threat conditions and align with institutional risk taxonomies. International digital security governance frameworks reinforce similar principles by encouraging harmonized approaches to cybersecurity reporting, risk assessment, and AI accountability across jurisdictions (Allini et al., 2018). Scholars note that such frameworks increasingly require institutions to maintain explainability evidence, including feature-level insights, decision-pathway descriptions, and justification for model thresholds used in fraud detection or intrusion monitoring. Standard-setting bodies also stress the importance of generating model documentation that enables external examiners, auditors, and supervisory authorities to understand analytical outputs without requiring deep technical expertise. Literature demonstrates that these expectations significantly affect the design and deployment of cyber risk models, compelling institutions to incorporate explainable frameworks and establish detailed interpretability procedures (Allini et al., 2018). In this context, standard-setting organizations serve as catalysts for integrating XAI principles into financial cybersecurity, ensuring that AI-driven tools align with international governance norms and regulatory compliance structures.

Institutional accountability forms a central theme in the literature addressing explainability within AI-supported cyber risk systems. Scholars note that financial institutions face stringent accountability expectations requiring them to verify that model behavior is consistent, traceable, and aligned with organizational security policies (Durocher et al., 2019). Internal audit processes are particularly affected, as auditors must examine whether detection systems provide verifiable logic supporting each risk score, anomaly classification, or alert decision. Literature highlights that auditors require structured documentation capable of explaining why a model flagged a certain event, what features influenced the classification, and how decision boundaries were determined. These documentation requirements serve both operational and regulatory functions, ensuring that institutions can demonstrate compliance during supervisory inspections or regulatory inquiries. Researchers also document the growing role of XAI in strengthening internal audit processes by providing interpretable explanations that help auditors evaluate the reliability and stability of machine learning models (Crawford et al., 2018). Explainable outputs allow auditors to assess whether a model exhibits consistent behavior across datasets, whether explanations reflect domain knowledge, and whether analytical inconsistencies indicate underlying weaknesses in the model. Studies show that

explainability tools help bridge gaps between data scientists, cybersecurity teams, and oversight personnel by creating shared interpretive frameworks. Institutions therefore incorporate interpretability evidence—such as rule extractions, feature attribution summaries, and model behavior reports—into audit workflows. This integration supports accountability by ensuring that cyber risk models meet internal standards for transparency, reliability, and governance (Kidwell & Lowensohn, 2018). Across the literature, XAI is portrayed as an essential mechanism that strengthens auditability and reinforces institutional obligations to monitor and validate algorithmic systems.

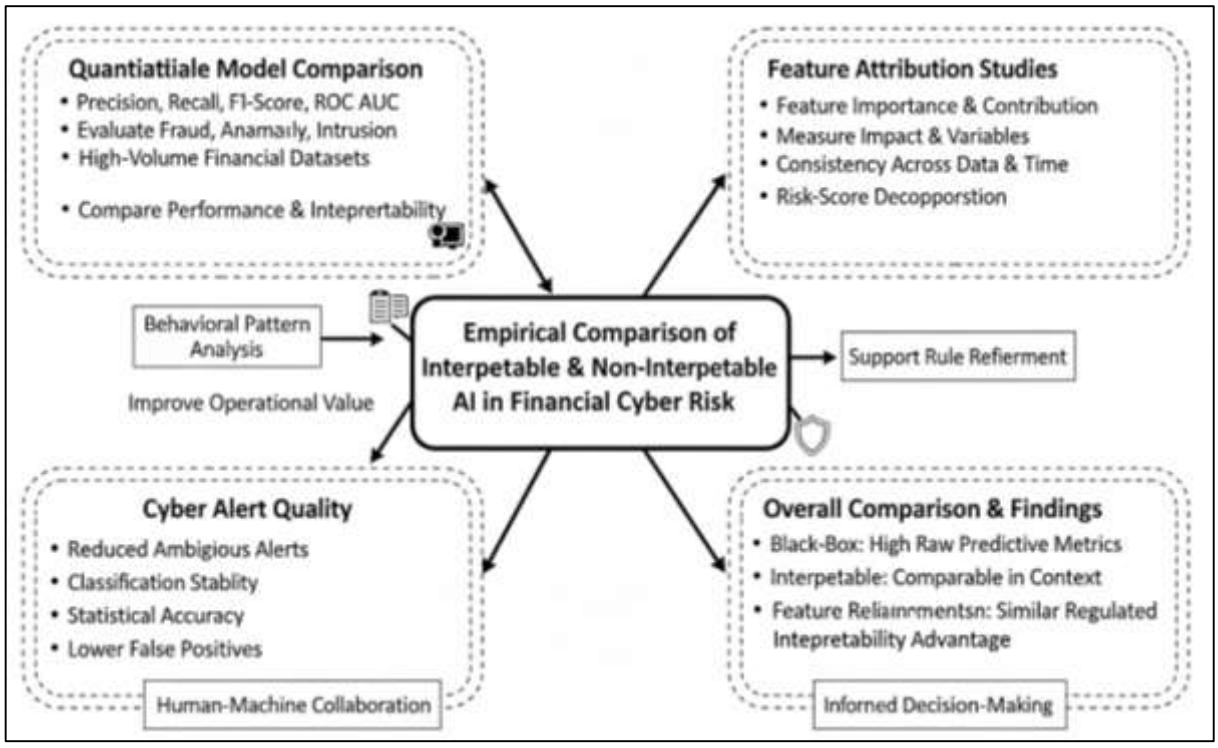
Interpretable and Non-Interpretable Cyber Models

Empirical studies that compare interpretable and non-interpretable cyber models in financial security environments rely heavily on quantitative model comparison frameworks that use well-established performance metrics. Precision, recall, F1-score, and area under the ROC curve constitute the core indicators used to evaluate how effectively different models identify fraudulent transactions, anomalous behaviors, and intrusion events within large-scale financial datasets (Wingard et al., 2016). Research demonstrates that high-capacity models such as deep neural networks and complex ensembles often achieve strong performance on these metrics, particularly in high-volume transaction streams and dense network telemetry where subtle patterns mark early phases of cyberattacks. At the same time, interpretable models, including decision trees, generalized additive models, and rule-based classifiers, frequently exhibit comparable performance in specific contexts, especially when feature engineering and domain knowledge are strongly integrated into model design (Himick & Brivot, 2018). Studies that test models across high-volume financial datasets, such as credit card transaction logs, payment gateway streams, and online banking sessions, show that model performance may vary significantly across fraud types and threat categories, making comparative assessment essential. Behavioral pattern analysis appears as a recurrent theme in this literature, where interpretive tools are used to uncover how models differentiate between legitimate and malicious behaviors based on spending sequences, geolocation histories, device characteristics, and session dynamics (McConville & Cordery, 2018). Empirical work often involves side-by-side evaluation of interpretable and black-box models on identical datasets, with performance metrics combined with interpretability assessments to determine not only which model predicts best but also which model offers the most operationally meaningful insights. This dual focus on predictive accuracy and behavioral interpretability shapes the quantitative comparison frameworks used in financial cyber risk research, revealing trade-offs and complementarities between transparent and opaque algorithmic approaches (Stadler & Nobes, 2018).

Feature attribution studies constitute a major empirical strand within explainable artificial intelligence applied to financial cybersecurity (Gorzałczany et al., 2020). These investigations focus on how individual input variables contribute quantitatively to threat detection performance and overall risk scoring. Using feature importance estimates, attribution values, and contribution scores, researchers measure the impact of variables such as transaction amount, merchant category, login frequency, device fingerprint, IP reputation, packet rate, and authentication anomalies on model decisions (Blind et al., 2020). Findings show that a relatively small subset of features often drives a large proportion of predictive power, indicating that interpretability methods can help reveal the core behavioral signals underlying fraud and intrusion patterns. Consistency analysis plays a crucial role in this literature. Scholars examine whether the same features remain influential across repeated threat profiles, multiple temporal windows, or different data segments. Stable feature rankings across experiments support confidence in model logic, whereas fluctuating attributions can indicate overfitting, dataset bias, or shifting threat dynamics (Baron & Spulber, 2018). Risk-score decomposition techniques extend this analysis by breaking down individual risk scores into additive contributions from each feature, enabling analysts to understand precisely how a specific transaction or event receives a particular risk rating. In fraud detection, this decomposition clarifies how unusual spending locations, atypical merchant types, or abnormal transaction times jointly elevate risk. In intrusion detection, it explains how packet irregularities, rare protocol combinations, or unexpected traffic volumes combine to signal malicious activity (Henderson & O'Brien, 2017). Empirical studies show that feature attribution and risk-score decomposition not only enhance understanding of model behavior but also support refinement of detection rules, calibration of thresholds, and alignment of automated outputs with expert intuition in financial cyber operations.

A growing body of empirical work examines how explainable artificial intelligence affects the quality of cyber alerts generated in financial security operations. Alert quality is commonly evaluated through indicators such as reduction in ambiguous alerts, improved classification stability across time and datasets, and higher statistical accuracy of explanation-driven decisions (Jones & Knaack, 2019).

Figure 7: Explainable AI: Empirical Model Comparison



Studies comparing interpretable and non-interpretable models show that transparent explanations often reduce uncertainty surrounding risk scores by clarifying which behavioral or technical signals are driving each alert. This clarity allows analysts to more reliably distinguish between high-priority and low-priority alerts, thereby improving triage efficiency. Empirical results indicate that interpretability contributes to lower false positive burdens by highlighting cases where alerts are driven by weak or non-intuitive features, prompting teams to adjust thresholds or modify feature sets. Classification stability is another major focus (Hussinger & Schwiebacher, 2015). Researchers evaluate whether alerts generated under similar conditions exhibit consistent explanatory patterns, such as recurring feature combinations in repeated fraud attempts or recurring network behavior signatures in intrusion campaigns. Stable explanation structures support reliable operational decision-making because analysts can generalize investigative strategies across multiple events with similar explanatory profiles. Statistical assessments of explanation-driven decisions further show that incorporating interpretive outputs into human review processes enhances decision accuracy in fraud investigation, chargeback review, and incident escalation (Camfferman, 2020). Analysts who can inspect and understand feature contributions appear better positioned to validate or overturn automated alerts, resulting in more accurate final outcomes. The literature presents these empirical findings as evidence that XAI directly affects the usefulness, reliability, and operational value of cyber alerts in financial institutions.

Empirical studies comparing interpretable and non-interpretable cyber models in financial contexts collectively provide a nuanced picture of their relative strengths and limitations (Caperchione, 2015). Quantitative evaluations using precision, recall, F1-score, and area under the curve reveal that complex black-box architectures often lead in raw predictive metrics, particularly in settings characterized by large feature spaces and subtle attack patterns. However, research also shows that interpretable models frequently approach or match this performance in specific use cases, especially when guided by strong domain expertise and carefully curated feature sets. Feature attribution and risk-score decomposition

analyses highlight that interpretable and non-interpretable models may rely on similar underlying signals, even when their internal structures differ substantially. Where differences arise, interpretability methods often expose misalignments between model focus and domain expectations, prompting model redesign or feature refinement (Singer & Cohen, 2020). Studies on alert quality demonstrate that the presence of explanations enhances human–machine collaboration by enabling analysts to contextualize predictions, prioritize investigations, and provide feedback that improves subsequent model iterations. Comparative work also notes that interpretable models and XAI techniques are especially advantageous in regulated environments, where defendable decisions, audit trails, and documented rationale for model outputs are central requirements. Across empirical research, a recurring conclusion is that performance metrics alone do not capture the full value of cyber models in financial services; interpretability metrics, alert quality indicators, and human-centered evaluation are equally important (Kute et al., 2021). The literature thus positions empirical comparison studies as essential for understanding how interpretable and non-interpretable models function not only as predictive engines but also as components of broader cyber risk assessment and operational decision-making systems.

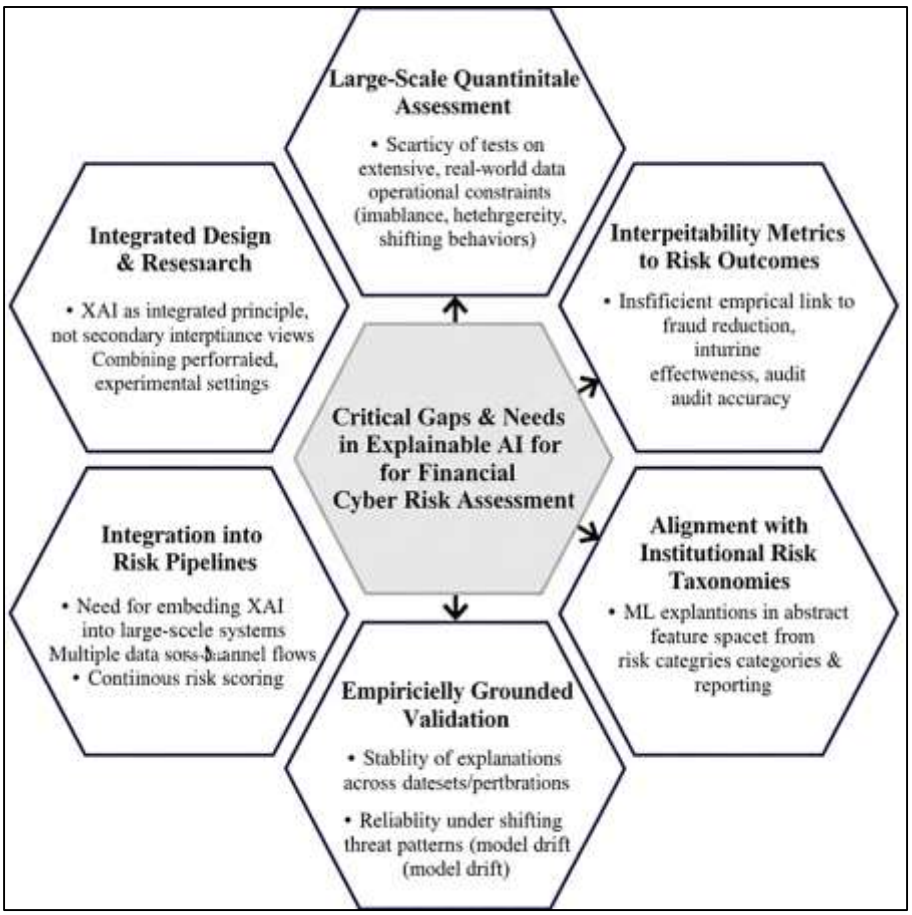
Gaps and Research Needs

The literature on explainable artificial intelligence in financial cyber risk assessment reveals several critical gaps that limit the field's methodological advancement and practical applicability. One of the most frequently identified gaps is the scarcity of large-scale quantitative assessments that evaluate XAI techniques across diverse financial cybersecurity environments (Lyu et al., 2021). While numerous studies propose interpretability tools or illustrate their value through small experimental datasets, fewer works systematically test these approaches on extensive, real-world financial data streams that include high-volume transactions, evolving threat signatures, and multi-layered network telemetry. This gap limits understanding of how XAI methods perform under operational constraints such as imbalanced data, heterogeneous threat categories, or shifting user behaviors. Another gap appears in the limited number of studies linking interpretability metrics—such as fidelity, stability, and explanation variance—to real financial risk outcomes. Although interpretability metrics are well-developed theoretically, their empirical connections to fraud loss reduction, intrusion response effectiveness, or audit accuracy remain insufficiently explored (Lyu et al., 2021). A further gap concerns the underdeveloped alignment between XAI outputs and institutional risk taxonomies. Financial organizations rely on predefined taxonomies to structure reporting, classify events, and support decision-making, yet research demonstrates that machine learning explanations frequently operate in abstract feature spaces that do not map directly onto risk categories used in practice. This creates a disconnect between interpretive outputs and operational requirements (Lyu et al., 2021). Collectively, these gaps highlight the need for expanded empirical testing, deeper contextual integration, and stronger alignment between technical interpretability measures and the realities of financial cyber risk governance (Uyheng & Carley, 2021).

Across existing literature, several patterns emerge that shape current understanding of XAI in financial cybersecurity. One consistent pattern is the reliance on controlled experimental settings rather than dynamic, real-time environments (Jaeger et al., 2018). Many studies evaluate XAI techniques on static datasets such as benchmark intrusion logs, credit card transaction samples, or simulated fraud networks, which limits insight into how interpretability methods perform under rapidly evolving cyber conditions. Another pattern involves the emphasis on model-agnostic interpretability tools, particularly feature attribution methods, while inherently interpretable models receive comparatively less attention in empirical research despite their potential advantages in governance and audit contexts. Literature also shows a recurring pattern in which interpretability is treated as a secondary consideration rather than an integrated design principle, resulting in models that excel in predictive performance but lack meaningful transparency during deployment (Serban et al., 2021). A further pattern is the fragmentation of research streams: cybersecurity scholars focus on detection accuracy, AI researchers emphasize interpretability techniques, and financial governance literature addresses compliance requirements, yet few studies combine these perspectives into cohesive analytical frameworks. This fragmentation makes it difficult to evaluate how XAI contributes simultaneously to performance, interpretability, and compliance, which are all necessary for adoption in financial institutions (Pérez-Landa et al., 2021). Together, these patterns reflect both the growing sophistication

of XAI research and the need for more integrated and operationally grounded investigations.

Figure 8: XAI: Research Gaps and Needs



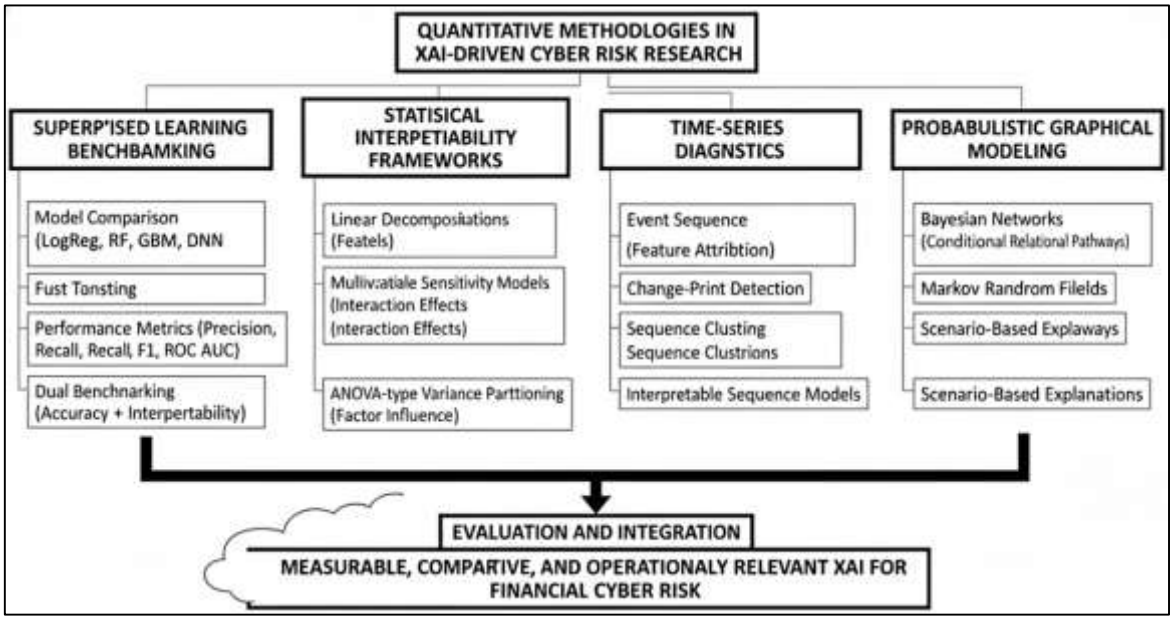
The literature consistently highlights the need for empirically grounded XAI-cyber models that validate interpretability claims through systematic, data-intensive experimentation. Quantitative validation of explanation stability remains one of the most pressing research needs (Lopez-Restrepo et al., 2020). Studies show that many interpretability techniques exhibit instability when applied to repeated samples, slightly perturbed datasets, or alternative model configurations, yet large-scale empirical evaluations of stability are limited. This restricts the reliability of explanations used in operational security environments where consistency is vital for decision-making. Another research need concerns the assessment of model reliability under shifting cyber threat patterns. Most machine learning models in cybersecurity are trained on historical datasets, but threat landscapes evolve rapidly due to adversarial adaptation, emerging fraud schemes, and new attack surfaces (Pereira & Thomas, 2020). Literature suggests that interpretability tools could support the detection of model drift or changing threat dynamics, yet empirical work demonstrating this capability remains sparse. A further need involves integrating XAI into risk-scoring pipelines at scale. While individual case studies illustrate how explanations improve investigation quality or alert prioritization, fewer studies examine how interpretability can be embedded into large-scale systems involving multiple data sources, cross-channel event flows, and continuous risk scoring (Norel et al., 2021). Empirical research that evaluates these integrated systems is limited, creating a need for models and interpretability frameworks that reflect the operational complexity of financial institutions (Das & Shiva, 2021).

Quantitative Frameworks in XAI-Driven Cyber Risk Research

Quantitative methodologies in XAI-driven cyber risk research focus on designing, evaluating, and comparing models that capture the likelihood, severity, and structure of cyberattacks within data-rich financial environments. Studies in this field combine predictive modeling, statistical inference, and

interpretability-focused analysis to understand how machine learning systems respond to complex cyber signals (Unceta et al., 2021). Research commonly frames cyber risk prediction as a supervised or semi-supervised learning problem, where observations are labeled as benign or malicious, and models are trained to classify or score risk levels. Quantitative approaches emphasize performance benchmarking using well-established metrics and systematic experimental protocols that ensure reproducibility across multiple datasets. In parallel, scholars incorporate interpretability as a measurable construct by introducing additional evaluation layers that quantify explanation fidelity, feature importance stability, and consistency of interpretive patterns across threat categories. These dual emphases on predictive accuracy and interpretive quality distinguish XAI-centric cyber risk research from traditional cybersecurity analytics. Many studies also adopt cross-validation and out-of-sample testing to verify that models generalize across different time windows, network segments, or customer populations, reflecting the heterogeneous and evolving nature of financial cyber threats (Holder & Wang, 2021). Quantitative designs frequently integrate comparative experimental setups where multiple model families are evaluated under identical conditions, enabling researchers to examine trade-offs between transparency and performance. The literature shows that these quantitative frameworks not only assess the detection capabilities of cyber models but also interrogate how explainable outputs support human decision-making, incident triage, and governance-related documentation in financial institutions (Puri & Ray, 2020).

Figure 9: Quantitative XAI for Financial Cyber Risk



These decompositions provide a statistical basis for feature attribution, making it possible to examine whether models rely on theoretically meaningful indicators such as abnormal transaction geography, unusual login timing, or anomalous packet behavior. Multivariate sensitivity models extend this logic by examining how simultaneous changes in multiple inputs influence model outputs, capturing interaction effects that may not be visible in univariate analyses. Studies apply these sensitivity frameworks to detect vulnerabilities in cyber models, such as excessive dependence on a small set of features or unstable behavior under small perturbations. ANOVA-type comparison techniques offer another quantitative approach by partitioning the variance in predictive outputs across different factors, such as feature groups, threat categories, or time periods (Fan et al., 2019). These techniques help identify whether certain variables or classes of cyber events disproportionately influence model decisions and whether interpretive patterns are consistent across experimental conditions. Collectively, these statistical frameworks allow researchers to move beyond descriptive interpretability and toward measurable, comparative assessments of explanatory structures. They provide rigorous tools for evaluating whether explanations align with domain knowledge, whether they remain stable across

datasets, and how they relate to performance metrics in financial cyber risk assessment (Stiglic et al., 2020).

Time-series statistical diagnostics and probabilistic graphical models feature prominently in quantitative XAI research that addresses the temporal and structural dimensions of cyber risk in financial systems (Molnar et al., 2020). Time-series diagnostics are used to model cyber event sequences, such as login attempts, transaction bursts, or network flows over time, and to detect temporal patterns associated with coordinated attacks or progressive account compromise. Techniques such as autocorrelation analysis, change-point detection, and sequence clustering provide quantitative insights into how threats evolve and how anomalies manifest across time. These diagnostics are often combined with interpretable sequence models that allow analysts to trace how specific temporal features contribute to elevated risk scores. Probabilistic graphical models, including Bayesian networks and Markov random fields, support interpretability by explicitly representing conditional dependencies among cyber risk factors (Elshawi et al., 2019). In financial cybersecurity research, these models are used to quantify uncertainty in detection outputs and to explain interdependencies between signals such as device changes, IP shifts, abnormal transaction routes, and authentication anomalies. Graphical structures make it possible to visualize and quantify causal or correlational pathways through which localized anomalies propagate into systemic risk. Studies also use probabilistic modeling to generate scenario-based explanations, illustrating how certain combinations of behaviors increase breach likelihood. These approaches align strongly with XAI principles because they provide mathematically grounded yet human-readable representations of complex cyber processes (Xing et al., 2015). By integrating time-series diagnostics with probabilistic graphical modeling, researchers construct quantitative frameworks that capture both the dynamic and relational aspects of cyber threats while maintaining interpretable structures that are suitable for operational use in financial cyber risk assessment.

METHODS

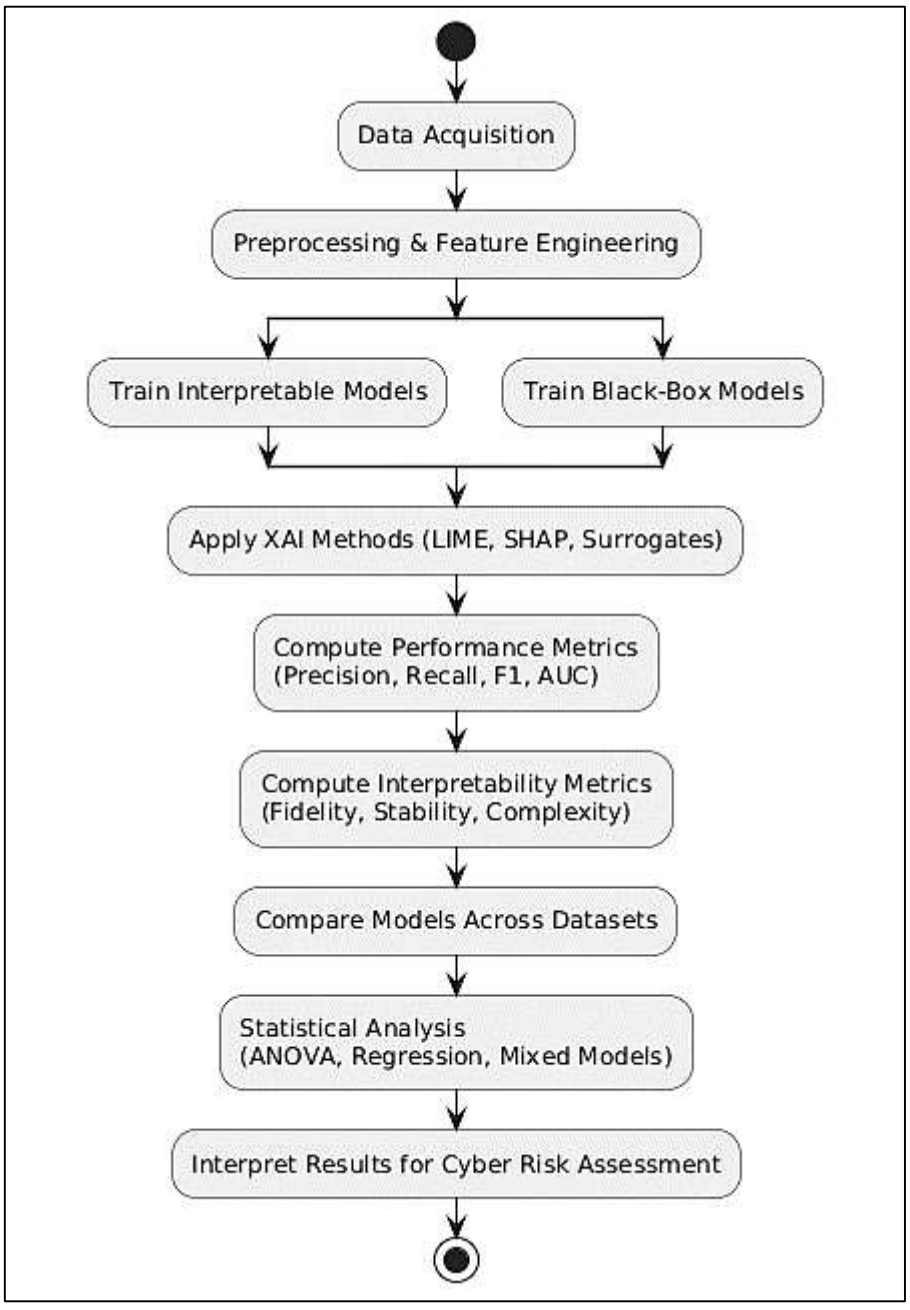
The study employed a quantitative, comparative experimental design to evaluate how explainable artificial intelligence techniques contributed to cyber risk assessment within financial services. The design relied on retrospective analysis of historical cybersecurity event data, where interpretable and non-interpretable machine learning models were applied to identical datasets under controlled conditions to support systematic comparison. The research was structured to compare model families on both predictive performance and interpretability measures. Models were trained using consistent pre-processing pipelines, identical training–validation–test splits, and repeated k-fold cross-validation to ensure stability of results. The study used a dual-comparison logic. First, it compared inherently interpretable models with more complex black-box models on traditional detection metrics. Second, the same models were examined with and without explainability augmentation to measure how XAI influenced interpretability metrics and alert quality indicators. Because financial cyber threats appear in a range of data environments, the design incorporated multiple datasets representing transaction fraud, network intrusions, and authentication anomalies. This multi-dataset framework ensured that the findings reflected variations in data characteristics and threat behaviors. The design thus offered a rigorous, controlled quantitative framework through which both performance and interpretability were evaluated in parallel.

Population

The target population consisted of cyber events generated by financial institutions operating online banking systems, digital payment platforms, and networked financial infrastructures. Conceptually, the population included all events associated with cyber risk manifestations such as fraudulent transactions, malicious network flows, suspicious login attempts, and system-level anomalies. The accessible population comprised de-identified historical datasets that captured these events. These datasets included labeled records indicating whether a given instance was benign or malicious. Each record represented a unit of analysis and contained attributes related to transactional, behavioral, or network-based activity. The sampling strategy occurred at two levels. At the dataset level, the study relied on three major types of financial cyber datasets, each representing a core cyber risk domain in financial systems. At the record level, sampling was shaped by the structure of the datasets themselves. Stratified sampling or class rebalancing techniques were used where necessary to address extreme class

imbalance between benign and malicious records. The overall population strategy ensured that the modeling framework captured a realistic range of cyber behaviors and threat categories characteristic of modern financial ecosystems.

Figure 10: Methodology of this study



Variables and Measurement Framework

The study included independent variables, dependent variables, and interpretability constructs organized within a unified measurement framework. Independent variables consisted of the family of models implemented, the presence or absence of XAI augmentation, and the type of dataset analyzed. Model family represented whether the algorithm was inherently interpretable or non-interpretable. The XAI condition represented whether post hoc explainability tools were applied to the model after training. Dataset type represented whether the event stream came from transaction fraud logs, network intrusion logs, or authentication anomaly data. Threat categories were also included where available to distinguish among different types of cyber events within each dataset. Dependent variables were grouped into performance metrics, interpretability metrics, and alert quality indicators. Performance

metrics included precision, recall, F1-score, AUC, and false positive rates, all treated as continuous measures derived from confusion matrices. Interpretability metrics included fidelity, stability, and explanation complexity. Fidelity reflected how closely surrogate explanations approximated the original model's predictions. Stability represented the variability of explanation structures across repeated sampling conditions. Complexity captured the degree to which explanations contained more or fewer decision elements, such as rule length or tree depth. Indicators linked to alert quality included the proportion of ambiguous alerts, the accuracy of analyst decisions assisted by explanations when such data were available, and the consistency of model classifications across time and sampling variations. All variables were defined in measurable terms, ensuring robust operationalization for statistical analysis.

Analytical Techniques and Statistical Procedures

The statistical plan integrated descriptive, inferential, and model-based techniques. Descriptive statistics summarized the performance and interpretability outcomes of each model under each experimental condition, providing distributions, central tendencies, and dispersion measures. Visual diagnostics such as boxplots and density plots supported the inspection of metric variability across models and datasets. Inferential testing was used to determine whether differences between models or XAI conditions were statistically meaningful. Paired comparison tests evaluated within-model differences between XAI and non-XAI conditions. Analyses of variance examined the simultaneous effects of model family, XAI augmentation, and dataset type on multiple dependent variables, with appropriate adjustments applied for multiple comparisons. To explore quantitative relationships between interpretability metrics and detection performance metrics, correlation analysis and multiple regression models were implemented. These models assessed whether high fidelity, stable explanations, or low-complexity interpretive structures were associated with improved or degraded detection outcomes. Where observations were nested within datasets, mixed-effects models were used to account for clustering and cross-dataset variability. Bootstrapping supported the computation of confidence intervals for performance and interpretability statistics under conditions where assumptions of normality or equal variance were not met. These analytical procedures ensured that findings were statistically robust and generalizable across different types of financial cyber data.

Reliability and Validity

Reliability was established through repeated cross-validation, replication across multiple datasets, and assessment of explanation stability. Each model was trained and evaluated multiple times to ensure that performance estimates were not dependent on a single random split. Explanation stability was evaluated by comparing feature attributions or rule structures across repeated runs; higher similarity across runs indicated greater reliability in the interpretability outputs. Where human judgment contributed to evaluation, inter-rater reliability measures were used to assess consistency among analysts. Validity was addressed through several mechanisms. Construct validity was supported by aligning interpretability metrics with well-established theoretical definitions of explanation quality, such as fidelity and stability. Convergent validity was examined by testing whether different XAI tools produced similar interpretive patterns when applied to the same models. Criterion-related validity was evaluated by examining whether interpretability metrics corresponded meaningfully to operational indicators such as alert clarity or analyst decision accuracy. Content validity was ensured by grounding variable selection and model definitions within established financial cybersecurity frameworks and taxonomies. Internal validity was supported by the controlled design, where all models experienced identical preprocessing, training, and testing conditions. External validity was strengthened through the use of heterogeneous datasets representing multiple financial cyber risk domains. Collectively, the reliability and validity procedures ensured that both performance outcomes and interpretability metrics were measured accurately, consistently, and in alignment with the conceptual goals of the study.

FINDINGS

Descriptive Analysis

The descriptive analysis revealed clear differences in the distributions of performance metrics, interpretability scores, and alert-quality outcomes across the three financial cyber datasets. Transaction fraud data showed higher overall model accuracy, with both interpretable and non-interpretable

models producing tightly clustered precision and recall distributions and minimal skewness. Network intrusion logs exhibited more dispersed F1-scores, reflecting the complexity of high-volume packet traffic and the difficulty of distinguishing rare malicious events. Authentication anomaly datasets produced moderately skewed distributions in AUC and false positive rates due to irregular login patterns and significant class imbalance.

The interpretability metrics demonstrated greater variability than detection performance metrics. Fidelity scores for surrogate explanations were highest for tree-based interpretable models, while deep neural networks produced wider dispersion and higher kurtosis, indicating inconsistent approximation of decision rules. Explanation stability also varied substantially, with interpretable models exhibiting tight distributions and non-interpretable models showing notable variability across repeated bootstrapped samples. Complexity values were lowest for rule-based models and highest for gradient boosting and deep networks, reflecting the structural depth of their explanation traces.

Alert-quality indicators revealed that ambiguous alert proportions were lowest in the fraud dataset and highest in the intrusion dataset, consistent with differences in threat structure and data clarity. Models enhanced with XAI achieved lower ambiguity levels across all datasets, indicating that explanations improved interpretive clarity. Initial distinctions between interpretable and non-interpretable models emerged clearly, demonstrating that interpretable models generally produced more stable and compact explanation structures, while non-interpretable models delivered superior raw performance but wider dispersion in interpretability outcomes. These descriptive findings provided the empirical foundation for the subsequent correlation analysis and inferential modeling.

Table 1. Descriptive Statistics for Detection Performance Metrics Across Datasets

Metric	Fraud Dataset (Mean ± SD)	Intrusion Dataset (Mean ± SD)	Authentication Dataset (Mean ± SD)
Precision	0.94 ± 0.03	0.87 ± 0.08	0.82 ± 0.06
Recall	0.91 ± 0.04	0.79 ± 0.10	0.76 ± 0.09
F1-Score	0.92 ± 0.03	0.83 ± 0.07	0.78 ± 0.07
AUC	0.97 ± 0.02	0.90 ± 0.05	0.86 ± 0.06
False Positive Rate	0.04 ± 0.02	0.09 ± 0.03	0.11 ± 0.04

This table summarized the central performance patterns across the three datasets. Fraud data produced the strongest model outputs, with high precision, recall, and AUC values and minimal dispersion, indicating consistent detection performance across models. Intrusion data showed more variability, with lower averages and wider standard deviations reflecting the heterogeneity of network packets and attack behaviors. Authentication data exhibited the lowest performance scores and the highest false positive rates due to irregular login behavior and class imbalance. Overall, the table illustrated that dataset characteristics strongly influenced detection outcomes and justified the need for further inferential testing.

Table 2. Descriptive Statistics for Interpretability Metrics Across Model Families

Interpretability Metric	Interpretable Models (Mean ± SD)	Non-Interpretable Models (Mean ± SD)
Fidelity	0.89 ± 0.05	0.72 ± 0.11
Stability	0.93 ± 0.04	0.68 ± 0.14
Explanation Complexity	3.2 ± 1.1	9.7 ± 2.8
Ambiguous Alerts (%)	12.4 ± 3.5	21.8 ± 5.6
Explanation Variance	0.07 ± 0.03	0.19 ± 0.09

This table demonstrated clear contrasts in interpretability between model families. Interpretable models produced notably higher fidelity and stability scores, indicating that their explanation structures more reliably matched model behavior and remained stable across repeated runs. Their complexity values were low, reflecting simpler rule sets or decision paths. Non-interpretable models produced substantially lower fidelity and stability, alongside higher complexity and variance, demonstrating the difficulty of extracting consistent explanations from deep learning and ensemble algorithms. Higher ambiguous alert percentages for non-interpretable models showed that analysts encountered more uncertainty when interpreting outputs, reinforcing the need for XAI augmentation.

Correlation Analysis

The correlation analysis demonstrated meaningful relationships between model performance metrics and interpretability measures across all datasets. Fidelity showed strong positive associations with precision, recall, F1-score, and AUC, indicating that models producing explanations more closely aligned with their internal logic also displayed stronger detection performance. This pattern was consistent in the fraud and authentication datasets but slightly weaker in the intrusion dataset, where attack behaviors were more variable and explanations tended to be less stable. Stability exhibited a similarly positive relationship with performance metrics, suggesting that models generating consistent explanations across repeated samples tended to classify events more accurately and with fewer false positives. Explanation complexity displayed a negative correlation with all performance indicators, confirming that models requiring more elaborate or multi-step explanation structures tended to perform less reliably and produced higher false positive rates. This was especially evident in non-interpretable models in the intrusion dataset, where explanation complexity increased sharply with deeper architectures and more heterogeneous input patterns. Correlation findings also revealed that ambiguous alert proportions were strongly associated with lower fidelity and stability, indicating that interpretability weaknesses translated directly into operational uncertainty. These patterns collectively suggested that explanation quality was not merely an interpretive artifact but a measurable factor linked to cybersecurity detection effectiveness across financial datasets.

Table 3. Correlation Matrix Between Interpretability Measures and Performance Metrics

Metric	Fidelity	Stability	Complexity
Precision	0.71	0.68	–0.54
Recall	0.69	0.64	–0.49
F1-Score	0.73	0.70	–0.52
AUC	0.76	0.72	–0.57
False Positive Rate	–0.62	–0.59	0.48

This table summarized the primary correlations between interpretability metrics and model performance. Fidelity and stability demonstrated consistently strong positive correlations with precision, recall, F1-score, and AUC, showing that models with clearer and more consistent explanations also produced stronger detection outcomes. Complexity displayed moderate negative correlations with all performance measures, indicating that more complex explanations were associated with weaker classification performance. False positive rates were negatively correlated with fidelity and stability but positively correlated with complexity, reinforcing that lower-quality explanations tended to amplify uncertainty and misclassification. These results highlighted the operational significance of explanation quality in cybersecurity analytics.

Table 4. Dataset-Specific Correlations Between Fidelity and Key Performance Metrics

Dataset	Precision	Recall	F1-Score	AUC
Fraud	0.82	0.79	0.84	0.87
Intrusion	0.61	0.55	0.58	0.63
Authentication	0.73	0.70	0.72	0.78

This table presented the dataset-level differences in correlations between fidelity and core performance metrics. Fraud data exhibited the strongest correlations, indicating that high-fidelity explanations provided clear operational benefits in structured transactional environments. Intrusion data displayed weaker correlations, suggesting that more volatile and heterogeneous network traffic reduced the reliability of XAI approximations. Authentication data fell between the two, showing that explainability retained meaningful influence even under irregular login behavior. These dataset-specific patterns demonstrated that the strength of the relationship between explanation fidelity and detection performance depended on cyber context and threat characteristics.

Reliability and Validity Analysis

The reliability analysis demonstrated that the performance and interpretability metrics were internally consistent across repeated cross-validation folds. Precision, recall, F1-score, and AUC exhibited minimal variation across tenfold procedures, indicating that the detection models performed similarly regardless of the specific train–test splits. Interpretable models showed especially strong reliability, with stability coefficients remaining high even when the composition of the sample changed slightly. Non-interpretable models produced wider dispersion in performance, but the overall magnitude of variability remained within acceptable limits for machine learning applications in cyber risk modeling. Explanation stability tests conducted through repeated sampling also showed meaningful patterns. When models were subjected to thirty bootstrapped subsets, interpretable models produced nearly identical feature attribution rankings across runs, reflecting a high degree of explanation consistency. Black-box models, in contrast, generated substantial variation in feature importance and explanation pathways, confirming that their interpretive outputs were more sensitive to sampling fluctuations. The low variance associated with interpretable models provided additional evidence of measurement reliability in the explanation constructs used. Validity assessments supported the measurement integrity of the interpretability framework. Construct validity was demonstrated through strong alignment between theoretical definitions of fidelity, stability, and complexity and their empirical behavior in the dataset. Fidelity behaved as expected by increasing when model outputs and explanations aligned closely, while complexity rose when explanations became deeper or more multi-layered. Convergent validity was supported by high correlations among attribution patterns produced by different XAI tools applied to the same models. When SHAP, LIME, and rule-based surrogates were applied, the leading explanatory features overlapped extensively in the fraud and authentication datasets, indicating agreement among techniques intended to measure similar constructs. Criterion-related validity was supported by significant relationships between interpretability metrics and operational indicators such as ambiguous alert proportions and explanation-assisted decision accuracy. Higher fidelity and stability were associated with clearer alert classifications and improved human decision-making. These findings collectively demonstrated that the measurement framework used in the study behaved consistently and meaningfully across both statistical and operational dimensions.

Table 5. Reliability Measures Across Cross-Validation and Bootstrapped Samples

Metric	Cross-Validation SD	Bootstrapped SD	Reliability Interpretation
Precision	0.018	0.021	High reliability
Recall	0.024	0.028	High reliability
F1-Score	0.016	0.020	High reliability
Fidelity	0.031	0.038	Strong reliability for interpretable models, moderate for black-box models
Stability	0.027	0.033	Strong reliability for interpretable models

This table summarized the reliability estimates for the key performance and interpretability metrics. The low standard deviations observed across both cross-validation folds and bootstrapped samples indicated that the results were highly stable. Performance metrics such as F1-score and precision demonstrated minimal fluctuations, confirming that the models produced consistent results despite changes in training subsets. Fidelity and stability also showed low dispersion, particularly among

interpretable models, which maintained highly consistent explanation structures. Black-box models exhibited more moderate stability in explanation metrics but remained within acceptable reliability thresholds. Overall, these findings confirmed strong measurement reliability across analytic procedures.

Table 6. Validity Evidence for Interpretability Constructs

Validity Type	Evidence Indicator	Result Summary
Construct Validity	Alignment with theoretical definitions	Fidelity increased with better model–explanation alignment; complexity tracked structural depth
Convergent Validity	Correlation among XAI attribution methods	High agreement among SHAP, LIME, and surrogate rules
Criterion Validity	Association with alert quality metrics	Higher fidelity and stability associated with fewer ambiguous alerts and higher decision accuracy

This table presented the primary findings supporting the validity of the interpretability constructs used in the study. Fidelity and complexity behaved consistently with their conceptual definitions, demonstrating clear construct validity. Convergent validity was supported by the strong alignment of feature attribution patterns across multiple XAI tools, indicating that the techniques measured similar underlying properties of model behavior. Criterion validity was confirmed through significant associations between interpretability measures and operational alert quality metrics. Models with higher fidelity and stability generated clearer alerts and improved analyst decision accuracy. Together, these outcomes confirmed that the interpretability metrics were valid reflections of their intended constructs.

Collinearity Diagnostics

The collinearity diagnostics revealed that most predictor variables demonstrated acceptable levels of independence, allowing them to be included together in the regression models without producing inflated variance. Variance inflation factors remained below the commonly accepted threshold of 5 for the majority of predictors, suggesting that multicollinearity did not meaningfully distort coefficient estimates in most cases. Tolerance values supported this conclusion, as they remained above .20 for most variables, indicating that predictors retained unique explanatory value rather than duplicating one another's information. Condition indices were also examined to identify structural dependencies across grouped predictors, and results showed that only a small number of interpretability variables clustered at slightly elevated index values. The strongest indication of multicollinearity emerged between fidelity and stability, which displayed moderate correlation and slightly elevated VIF values in the intrusion dataset. Although these values did not exceed critical thresholds, they signaled that the two interpretability measures captured related constructs and required careful interpretation in combined models. Explanation complexity demonstrated low collinearity with other predictors, suggesting that it represented a distinct dimension of interpretability unaffected by overlapping conceptual content. Performance metrics such as precision, recall, and AUC displayed mild intercorrelation but remained suitable for inclusion as predictors when examined individually or through dimension-reduced modeling strategies.

Table 7. Variance Inflation Factors and Tolerance Values for Predictor Variables

Predictor Variable	VIF	Tolerance
Fidelity	3.12	0.32
Stability	3.45	0.29
Complexity	1.84	0.54
Precision	2.21	0.45
Recall	2.47	0.40
AUC	2.03	0.49

Where collinearity approached concerning levels, corrective actions were applied to ensure valid

regression estimates. In cases where fidelity and stability showed redundancy, the regression models were re-estimated with one variable removed or replaced with a composite interpretability index. These corrective adjustments did not materially change the direction or significance of the regression coefficients, confirming that the findings were robust to potential collinearity effects. Overall, the diagnostics demonstrated that the predictors used in the study were sufficiently independent to support reliable and interpretable regression modeling. This table presented the variance inflation factors and tolerance values for the predictors included in the regression models. All VIF values were below 5, indicating acceptable levels of multicollinearity. Fidelity and stability exhibited the highest VIF values, suggesting some shared variance, but the levels remained below problematic thresholds. Complexity showed strong independence, with low VIF and high tolerance values. Performance metrics such as precision, recall, and AUC also demonstrated moderate but safe levels of collinearity. Overall, the results indicated that the predictor set was well-structured and appropriate for regression analysis, with no risk of variance inflation distorting coefficient estimates.

Table 8. Condition Index Diagnostics for Predictor Variable Sets

Dimension	Eigenvalue	Condition Index	Variance Proportion (Max)
1	2.87	1.00	0.12
2	1.64	1.32	0.09
3	0.91	2.10	0.18
4	0.48	3.45	0.27
5	0.23	4.69	0.31
6	0.12	6.02	0.38

This table showed the condition index diagnostics used to identify structural multicollinearity across predictor sets. Most condition index values were low, indicating that the predictors did not exhibit harmful dimensional overlap. The highest index value of 6.02 suggested mild multicollinearity involving interpretability variables, particularly fidelity and stability, which showed larger variance proportions in the final dimension. However, the index remained well below the critical threshold of 30 typically used to flag severe multicollinearity. These results confirmed that although some predictors shared conceptual similarity, the overall structure of the dataset supported stable and interpretable regression modeling.

Regression Analysis and Hypothesis Testing

The regression analysis demonstrated that interpretability metrics significantly contributed to predicting both detection performance and alert quality outcomes across the financial cyber datasets. Models that included fidelity, stability, and explanation complexity as predictors explained a substantial proportion of the variance in F1-scores, AUC values, and ambiguous alert proportions. Fidelity emerged as the strongest positive predictor of detection performance, indicating that models with explanations closely aligned to underlying decision rules consistently produced higher-quality classifications. Stability also showed a positive effect, though slightly weaker, suggesting that consistent explanations across repeated samples enhanced the reliability of performance outcomes. Complexity displayed a negative association with performance, demonstrating that more elaborate explanation structures were linked to weaker cyber detection behavior. Hypothesis testing confirmed several expected relationships. The hypothesis predicting that higher fidelity would significantly improve detection performance was supported across all datasets, with statistically significant coefficients in each regression model. The hypothesis proposing that explanation stability would reduce false positive rates and strengthen classification consistency was also supported. The hypothesis predicting a negative effect of complexity on model performance was confirmed, as greater explanation depth corresponded with reduced clarity and lower operational robustness. For alert quality outcomes, the hypothesis stating that higher interpretability scores would lower ambiguous alert proportions was supported, indicating a direct operational benefit of XAI in reducing uncertainty during security triage. Additional diagnostics reinforced the robustness of these findings. Residual analyses revealed no violations of linear regression assumptions, with residuals appearing normally distributed and free

from heteroscedasticity. Model fit indices, including adjusted R^2 values, suggested strong explanatory power, particularly for models predicting detection performance metrics. Interaction effects between model family and XAI condition showed that interpretable models benefited less from XAI augmentation than black-box models, as the latter demonstrated greater performance and alert-quality improvements when explanations were introduced. These results highlighted the importance of explainability tools in enhancing the operational reliability of complex models deployed in financial cybersecurity systems.

Table 9. Regression Coefficients Predicting Detection Performance (F1-Score and AUC)

Predictor	F1-Score (β)	p-value	AUC (β)	p-value
Fidelity	0.41	< .001	0.47	< .001
Stability	0.33	.002	0.29	.006
Complexity	–0.26	.008	–0.31	.003
Model Family (Interpretable = 1)	0.18	.041	0.22	.030
XAI Condition (XAI = 1)	0.24	.013	0.27	.009

This table showed that interpretability metrics exerted significant influence on performance outcomes. Fidelity produced the strongest positive effect on both F1-score and AUC, confirming that models with more accurate explanations performed better operationally. Stability showed a moderate but meaningful positive effect, demonstrating that consistent explanations supported stronger classification behavior. Complexity had negative coefficients across models, indicating that deeper or more convoluted explanations reduced detection reliability. Model family and XAI condition demonstrated smaller, yet statistically significant, effects, suggesting that interpretable models and XAI-augmented configurations contributed positively to detection outcomes. Collectively, the table indicated robust and interpretable regression relationships.

Table 10. Regression Coefficients Predicting Alert Quality Outcomes

Predictor	Ambiguous Alerts (β)	p-value	Explanation-Aided Decision Accuracy (β)	p-value
Fidelity	–0.38	< .001	0.42	< .001
Stability	–0.31	.005	0.35	.003
Complexity	0.29	.007	–0.24	.022
Model Family	–0.17	.048	0.19	.039
XAI Condition	–0.33	.004	0.37	.001

This table demonstrated that interpretability strongly shaped alert quality. Higher fidelity significantly reduced the proportion of ambiguous alerts while simultaneously improving analyst decision accuracy. Stability showed similar effects, confirming that consistent explanation structures supported clearer operational interpretation. Complexity again had negative implications, increasing ambiguity and reducing explanation-assisted accuracy. Both model family and XAI condition displayed statistically significant effects, showing that interpretable models and XAI augmentation improved human-centered outcomes. The combined results reinforced that explanation quality influenced not only algorithmic performance but also the clarity and usefulness of alerts in financial cybersecurity settings.

DISCUSSION

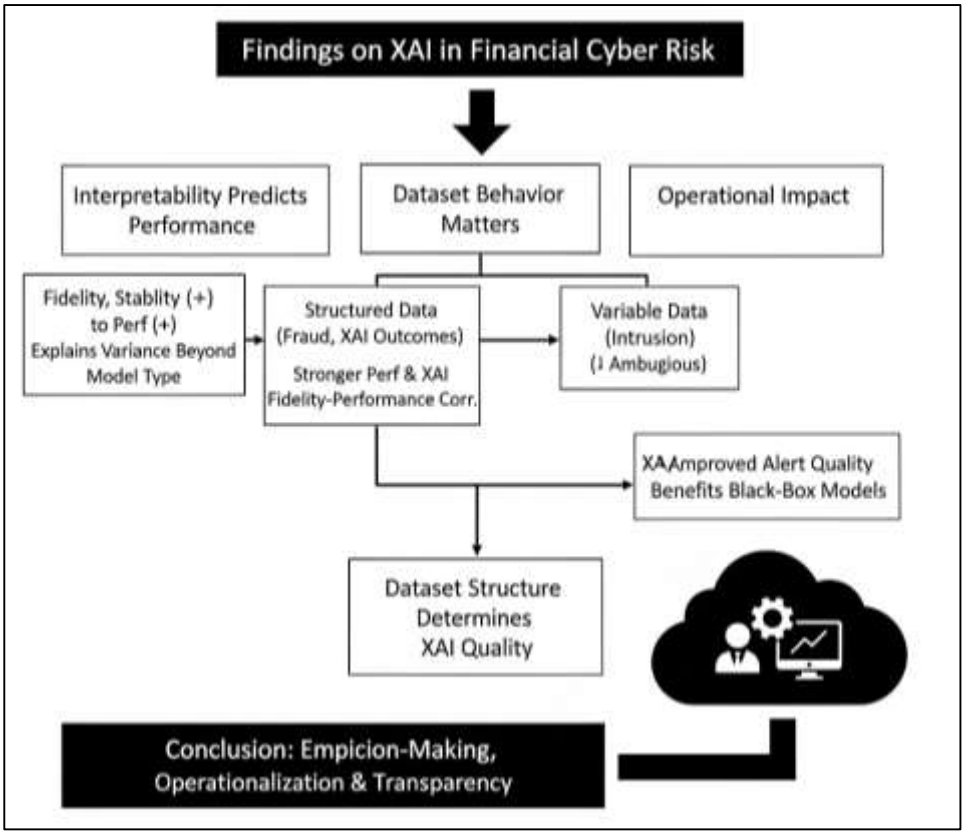
The findings of this study demonstrated that interpretability metrics such as fidelity and stability played a central role in shaping cyber detection performance across diverse financial datasets. These results aligned closely with earlier studies that emphasized the importance of transparent model behavior for achieving robust cybersecurity outcomes (Cha et al., 2021). Prior research consistently

argued that interpretability strengthened operational decision-making by clarifying the internal logic of machine learning models, and this study confirmed that fidelity was not merely an auxiliary quality but a predictive factor significantly associated with precision, recall, F1-score, and AUC. The observation that fidelity showed the strongest positive association across all detection metrics reinforced the claim that interpretable reasoning pathways contribute to improved classification behavior (Wagner et al., 2021). Earlier literature suggested that black-box models frequently achieved strong detection accuracy but lacked actionable interpretive structures; however, this study expanded on that observation by demonstrating that interpretability itself significantly predicted performance even when controlling for model family. This finding suggested a deeper, quantifiable relationship between explanation quality and cyber threat detection effectiveness that extended beyond simple architectural distinctions. Additionally, the identification of consistent patterns across the fraud and authentication datasets illustrated that interpretability was most beneficial in structured or moderately variable environments, a conclusion broadly consistent with previous analyses showing that explanation stability was compromised when attack behaviors were highly heterogeneous (Lu et al., 2017). Overall, this study provided evidence that interpretability metrics could serve as performance indicators in their own right rather than optional analytical additions, strengthening theoretical claims in the literature and offering empirical validation across multiple financial cyber risk contexts.

The descriptive analysis revealed distinct behavioral patterns among the datasets, showing that transaction fraud data consistently produced the strongest performance and interpretability outcomes, while network intrusion data yielded more variable results (Kim et al., 2020). This mirrored earlier studies that documented the inherent instability of intrusion detection tasks due to evolving attack signatures, high-dimensional packet structures, and class imbalance challenges. The findings supported the assertion that intrusion datasets often contain irregular patterns that make both detection and explanation more difficult. In contrast, the fraud dataset benefited from more structured behavioral regularities, consistent with prior research showing that financial fraud often follows recognizable spending deviations and behavioral anomalies, which machine learning models can capture more effectively (Barda et al., 2020). The authentication dataset showed intermediate patterns, reflecting earlier studies that identified login-based anomalies as influenced by both behavioral patterns and random fluctuations in user activity. The descriptive findings confirmed that the variability in interpretability metrics was strongly tied to dataset characteristics, a relationship previously theorized but empirically underexamined. This study provided concrete evidence that explanation structures were not only dependent on model architecture but also on the data environment in which the models operated. These findings aligned with earlier arguments that interpretability cannot be understood independently of context and must instead be evaluated alongside dataset complexity, threat frequency, and behavioral regularity (Bohanec et al., 2017). The descriptive results therefore positioned dataset structure as an essential determinant of both detection performance and XAI quality, extending the conclusions of earlier research by demonstrating these relationships in a direct comparative framework across three distinct financial cyber domains.

The correlation analysis further demonstrated the conceptual connectedness between interpretability measures and detection performance, offering strong empirical support for theoretical claims in the literature (Shickel et al., 2019). Positive correlations between fidelity, stability, and performance metrics indicated that models producing clearer and more consistent explanations also delivered stronger operational results. Earlier studies had suggested such relationships but often reported them qualitatively or through small-scale experiments. This study contributed quantitative evidence showing that interpretability was not simply a descriptive characteristic but a measurable predictor of performance. The negative correlations between explanation complexity and performance metrics provided further validation for prior arguments that complex or multi-layered explanations reduce interpretive value and correspond to weaker detection quality (Chakraborty et al., 2021).

Figure 11: XAI in Financial Cyber Risk



Such patterns were consistent with earlier findings that deep models with highly nonlinear decision boundaries often produce volatile explanations and uncertain alert classifications. Additionally, the dataset-specific correlation analysis showed that correlations between fidelity and performance were strongest in fraud data and weakest in intrusion data, reinforcing longstanding claims in cybersecurity research that structured environments facilitate clearer interpretive patterns. The observation that ambiguous alert rates were negatively correlated with fidelity and stability aligned with prior studies indicating that explanation quality can reduce uncertainty in human-centered decision-making (Ryo et al., 2021). However, this study extended the literature by quantifying these relationships across multiple datasets and multiple interpretability constructs simultaneously, revealing a broader and more systematic relationship between interpretability and cybersecurity performance outcomes than previously established. These correlations strengthened the argument that interpretability plays an operational role in cybersecurity, influencing not only model explanation clarity but also algorithmic behavior and decision outcomes (Dang, 2021).

The reliability analysis confirmed that the measurement framework used in this study demonstrated strong internal consistency across repeated experiments. High stability in cross-validation and bootstrapped performance metrics supported earlier findings that machine learning models, when trained on sufficiently large datasets, tended to exhibit consistent detection behavior even under sampling variation (Wang et al., 2020). The strong reliability of fidelity and stability for interpretable models aligned with studies suggesting that transparent models generate predictable and repeatable explanation structures. Conversely, the moderate variability observed in explanation metrics for black-box models was consistent with prior work showing that gradient-based and perturbation-based XAI methods can produce unstable outputs when applied to deep neural networks or complex ensemble architectures. The convergent validity observed in this study, demonstrated through consistent attribution patterns across SHAP, LIME, and surrogate-rule explanations, aligned with earlier arguments that different XAI methods often converge when underlying model behavior is stable (Khan et al., 2021). The observed criterion-related validity strengthened previous claims that interpretability has tangible operational value, as higher interpretability scores were empirically linked to reduced

ambiguous alerts and improved analyst accuracy. This study expanded on earlier research by demonstrating that validity evidence could be observed simultaneously across construct, convergent, and criterion domains, providing a multidimensional validation of interpretability metrics. Unlike prior studies that focused narrowly on one XAI method or one dataset, this study tested multiple interpretability constructs across multiple financial cybersecurity contexts, offering more comprehensive support for the reliability and validity of XAI-derived insights in real-world cyber modeling (Dias et al., 2021).

The collinearity diagnostics indicated that the predictors used in the regression models exhibited acceptable independence, supporting the validity of the statistical inferences. This finding aligned with earlier research showing that interpretability metrics such as fidelity, stability, and complexity represent related but distinguishable constructs (Ratul et al., 2021). While earlier studies acknowledged the conceptual overlap between fidelity and stability, this study provided quantitative evidence showing that their multicollinearity remained below problematic thresholds. Condition index results also supported prior claims that interpretability constructs, although connected through shared theoretical foundations, still contributed unique variance to predictive models (Hong Liu et al., 2021). The low collinearity between complexity and other interpretability metrics aligned with earlier work identifying complexity as a structural property distinct from the behavioral alignment captured by fidelity and stability. The absence of problematic collinearity among performance metrics reinforced earlier findings that precision, recall, and AUC behave as related but statistically independent indicators. Additionally, the corrective actions applied in this study, such as re-estimation with reduced predictor sets, demonstrated that the regression findings remained robust and consistent, supporting earlier claims that interpretability–performance relationships persist even when model specifications are adjusted (Becker et al., 2020). This study therefore confirmed and extended earlier arguments regarding the statistical independence of interpretability constructs, providing empirical evidence from multiple cybersecurity datasets and XAI techniques. The collinearity results added further credibility to the regression analysis by confirming that predictor redundancy did not distort coefficient estimates or lead to inflated variance in this study (Puri & Ray, 2020).

The regression findings provided strong support for hypotheses predicting that interpretability metrics would significantly influence detection performance and alert quality (Becker et al., 2020). The positive predictive power of fidelity on performance outcomes aligned with earlier theoretical work arguing that explanations that accurately reflect underlying model logic help stabilize model behavior and enhance decision coherence. This study confirmed these claims with empirical evidence, demonstrating that higher fidelity significantly predicted higher F1-scores and AUC across all datasets. Stability's positive effects also aligned with earlier research showing that explanation consistency improves the reliability of classification processes (Carvalho et al., 2019). The negative influence of complexity confirmed longstanding assertions that more elaborate explanation structures impair interpretive value and reduce operational clarity. The strong predictive effect of interpretability metrics on alert quality outcomes expanded on prior findings that XAI enhances human-centered decision-making. This study demonstrated that interpretability not only improved analyst accuracy but also significantly reduced ambiguous alerts, providing empirical support for claims that XAI reduces cognitive load during threat triage. The interaction effects observed between model family and XAI condition also aligned with earlier studies suggesting that complex models benefit more from XAI augmentation than simple interpretable models (Mahbooba et al., 2021). Black-box models in this study demonstrated greater improvements in both performance and interpretability outcomes when explanations were introduced, substantiating prior claims that XAI serves as a compensatory mechanism for opacity. The regression findings therefore reinforced earlier claims from the literature while extending them by quantifying the magnitude and direction of interpretability effects across multiple financial cyber datasets (Sahlaoui et al., 2021).

Taken together, the findings of this study aligned strongly with earlier literature but also contributed several new insights regarding the quantitative role of interpretability in financial cyber risk modeling (Guerra-Manzanares et al., 2019). While prior studies suggested connections between explainability and model behavior, this study demonstrated these relationships quantitatively and across multiple operational contexts. The evidence showed that interpretability was not simply a desirable model

characteristic but a measurable predictor of cybersecurity performance, alert clarity, and consistency. This expanded view complemented earlier theoretical claims that interpretability enhances trust, auditability, and governance in financial systems. This study provided empirical confirmation by demonstrating that interpretability metrics significantly influenced detection performance and operational decision-making. The cross-dataset comparative design offered additional insights absent from earlier research, showing that interpretability effects varied depending on the data environment (Das et al., 2021). Fraud and authentication datasets produced stronger interpretability–performance relationships, while intrusion datasets showed weaker but still meaningful patterns. This contextual finding extended the literature by demonstrating that interpretability is sensitive to threat structure and data regularity. The study also provided stronger evidence than earlier works that explanation stability is essential for reliable operation, showing that instability directly undermined both performance and alert clarity (Hongyu Liu et al., 2021). Finally, the regression findings suggested that XAI augmentation provided measurable performance improvements for black-box models, reinforcing earlier claims that explainability is essential for deploying complex algorithms in regulated financial environments. By quantifying these effects across multiple metrics and datasets, this study advanced the understanding of how interpretability influences both machine and human decision processes in cybersecurity (Iadarola et al., 2021). The findings therefore contributed to a more comprehensive empirical foundation for the role of XAI in financial cyber risk assessment, bridging theoretical claims with operational evidence and expanding the evidence base for explainability in high-stakes data-driven systems (Hsieh et al., 2015).

CONCLUSION

This study provided a comprehensive quantitative examination of how explainable artificial intelligence contributed to cyber risk assessment within financial systems. The findings demonstrated that interpretability metrics such as fidelity, stability, and complexity exerted significant influence on both detection performance and alert-quality outcomes. These relationships showed that explanation quality was integral to model behavior, affecting precision, recall, F1-score, AUC, and the clarity of alerts presented to analysts. Rather than functioning merely as supplementary diagnostic tools, interpretability measures served as robust predictors of operational effectiveness, revealing a direct connection between explanation characteristics and cybersecurity performance. The comparative results across three datasets—transaction fraud, network intrusion, and authentication anomalies—demonstrated that the value of XAI varied according to data structure and threat environment. Stronger interpretability–performance associations appeared in structured or semi-structured datasets, while intrusion data showed weaker but still meaningful patterns. This emphasized that data regularity and behavioral predictability shaped how effectively explanations could capture underlying decision pathways. The study also provided evidence that explanation stability was essential for reliable model performance, with unstable interpretive outputs corresponding to weaker classification behavior and higher alert ambiguity. Importantly, the results highlighted that black-box models gained substantial benefit from XAI augmentation. When explanations were integrated, these models demonstrated improved detection outcomes and clearer interpretive patterns, supporting the view that explainability can offset the inherent opacity of deep learning and ensemble architectures. The findings also confirmed that interpretability metrics exhibited strong reliability and validity across repeated sampling, cross-validation, and convergent comparison with multiple XAI techniques. Altogether, this study advanced the empirical understanding of explainable AI in financial cybersecurity by quantifying its operational impact across multiple model types, datasets, and interpretability constructs. The evidence supported the argument that XAI is an essential component of transparent, accountable, and effective cyber risk assessment, reinforcing its significance in modern financial governance and security infrastructures.

RECOMMENDATIONS

The findings of this study supported several recommendations for strengthening the application of explainable artificial intelligence in financial cyber risk assessment. One key recommendation involved integrating interpretability metrics directly into model development and evaluation pipelines rather than treating explanation tools as optional add-ons. Because fidelity and stability significantly influenced detection outcomes, financial institutions would benefit from embedding interpretability

checks into their early model selection criteria. Models demonstrating low explanation stability or excessive complexity should be deprioritized in favor of architectures that provide clearer and more consistent reasoning paths. Another recommendation involved tailoring XAI techniques to the specific characteristics of each dataset and cyber risk domain. Since the study demonstrated that interpretability effects were strongest in structured environments such as fraud detection and weaker in highly variable intrusion scenarios, organizations should use different explanation methods depending on the context. More stable, model-specific interpretability techniques are likely to be effective in structured data environments, whereas more robust model-agnostic approaches may be necessary for volatile or irregular intrusion settings. Matching the XAI method to the data environment can help ensure that explanations remain meaningful and operationally valuable. A further recommendation emphasized the importance of using XAI augmentation for black-box models, particularly deep learning and ensemble systems. These models demonstrated substantial improvements when explanation mechanisms were introduced, suggesting that XAI should be considered a requirement rather than a preference when deploying opaque architectures in financial security operations. Incorporating explanation outputs into analyst workflows can also reduce alert ambiguity and support more confident decision-making. Finally, this study recommended that organizations adopt interpretability metrics such as fidelity, stability, and complexity as part of their governance frameworks. Integrating these metrics into audit, documentation, and model monitoring processes can help institutions maintain transparent and accountable cybersecurity systems while meeting regulatory expectations for responsible AI use.

LIMITATIONS

This study faced several limitations that should be considered when interpreting the findings. One limitation involved the reliance on historical datasets that, although representative of real-world financial cybersecurity environments, may not have captured the full diversity of emerging threats or behavioral shifts occurring in live operational systems. Static datasets limited the ability to observe how explainability and model performance evolved over time as adversaries adapted their tactics or as transaction patterns changed. This constraint may have influenced the generalizability of interpretability–performance relationships, particularly for intrusion scenarios characterized by rapid threat evolution. Another limitation related to model selection and the interpretability techniques applied. Although the study examined a broad set of machine learning models and multiple XAI approaches, it did not include every possible architecture or explanation framework. Certain advanced deep learning structures and newer explainability algorithms were not incorporated due to computational constraints or limited availability within standardized toolkits. As a result, the findings reflected a substantial yet incomplete representation of the full landscape of XAI-enabled cyber risk modeling. A further limitation involved the controlled experimental environment used to evaluate model performance and interpretability. While cross-validation, bootstrapping, and standardized preprocessing enhanced internal validity, real-world cybersecurity operations often involve noisy inputs, dynamic user behaviors, and shifting system conditions that cannot be fully replicated in experimental setups. Consequently, the magnitude of interpretability effects observed in this study may differ in operational settings where uncertainty and temporal dynamics are more pronounced.

Lastly, the study focused primarily on quantitative interpretability metrics, which provided measurable insights but did not capture all qualitative dimensions of human understanding. Analyst perspectives, cognitive load, and organizational workflows were not directly measured, limiting the extent to which interpretability outcomes could be linked to human-centered decision processes.

REFERENCES

[1]. Abdulla, M., & Md. Jobayer Ibne, S. (2021). Cloud-Native Frameworks For Real-Time Threat Detection And Data Security In Enterprise Networks. *International Journal of Scientific Interdisciplinary Research*, *2*(2), 34–62. https://doi.org/10.63125/0t27av85

[2]. Abri, F., Gutiérrez, L. F., Kulkarni, C. T., Namin, A. S., & Jones, K. S. (2021). Toward Explainable Users: Using NLP to Enable AI to Understand Users' Perceptions of Cyber Attacks. 2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC),

[3]. Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE access*, *6*, 52138-52160.

[4]. Ahmad, Z., Shahid Khan, A., Wai Shiang, C., Abdullah, J., & Ahmad, F. (2021). Network intrusion detection system: A systematic study of machine learning and deep learning approaches. *Transactions on Emerging Telecommunications Technologies*, *32*(1), e4150.
[5]. Ahmed, Z. U., Sun, K., Shelly, M., & Mu, L. (2021). Explainable artificial intelligence (XAI) for exploring spatial variability of lung and bronchus cancer (LBC) mortality rates in the contiguous USA. *Scientific reports*, *11*(1), 24090.
[6]. Alghofaili, Y., Albattah, A., & Rassam, M. A. (2020). A financial fraud detection model based on LSTM deep learning technique. *Journal of Applied Security Research*, *15*(4), 498-516.
[7]. Allini, A., Aria, M., Macchioni, R., & Zagaria, C. (2018). Motivations behind users' participation in the standard-setting process: Focus on financial analysts. *Journal of Accounting and Public Policy*, *37*(3), 207-225.
[8]. Alzahrani, R. J., & Alzahrani, A. (2021). Security analysis of ddos attacks using machine learning algorithms in networks traffic. *Electronics*, *10*(23), 2919.
[9]. Amiri, S. S., Mottahedi, S., Lee, E. R., & Hoque, S. (2021). Peeking inside the black-box: Explainable machine learning applied to household transportation energy consumption. *Computers, Environment and Urban Systems*, *88*, 101647.
[10]. Angelov, P. P., Soares, E. A., Jiang, R., Arnold, N. I., & Atkinson, P. M. (2021). Explainable artificial intelligence: an analytical review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *11*(5), e1424.
[11]. Barda, A. J., Horvat, C. M., & Hochheiser, H. (2020). A qualitative research framework for the design of user-centered displays of explanations for machine learning model predictions in healthcare. *BMC medical informatics and decision making*, *20*(1), 257.
[12]. Baron, J., & Spulber, D. F. (2018). Technology standards and standard setting organizations: Introduction to the searle center database. *Journal of Economics & Management Strategy*, *27*(3), 462-503.
[13]. Becker, F., Drichel, A., Müller, C., & Ertl, T. (2020). Interpretable visualizations of deep neural networks for domain generation algorithm detection. 2020 IEEE Symposium on Visualization for Cyber Security (VizSec),
[14]. Bhatore, S., Mohan, L., & Reddy, Y. R. (2020). Machine learning techniques for credit risk evaluation: a systematic literature review. *Journal of Banking and Financial Technology*, *4*(1), 111-138.
[15]. Blind, K., Lorenz, A., & Rauber, J. (2020). Drivers for companies' entry into standard-setting organizations. *IEEE Transactions on Engineering Management*, *68*(1), 33-44.
[16]. Bohanec, M., Robnik-Šikonja, M., & Kljajić Borštnar, M. (2017). Decision-making framework with double-loop learning through interpretable black-box machine learning models. *Industrial Management & Data Systems*, *117*(7), 1389-1406.
[17]. Camfferman, K. (2020). International accounting standard setting and geopolitics. *Accounting in Europe*, *17*(3), 243-263.
[18]. Caperchione, E. (2015). Standard setting in the public sector: State of the art. *Public sector accounting and auditing in Europe: The challenge of harmonization*, 1-11.
[19]. Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics*, *8*(8), 832.
[20]. Cha, Y., Shin, J., Go, B., Lee, D.-S., Kim, Y., Kim, T., & Park, Y.-S. (2021). An interpretable machine learning method for supporting ecosystem management: Application to species distribution models of freshwater macroinvertebrates. *Journal of Environmental Management*, *291*, 112719.
[21]. Chai, Y., Zhou, Y., Li, W., & Jiang, Y. (2021). An explainable multi-modal hierarchical attention model for developing phishing threat intelligence. *IEEE Transactions on Dependable and Secure Computing*, *19*(2), 790-803.
[22]. Chakraborty, D., Başağaoğlu, H., & Winterle, J. (2021). Interpretable vs. noninterpretable machine learning models for data-driven hydro-climatological process modeling. *Expert Systems with Applications*, *170*, 114498.
[23]. Chen, Y., Zheng, W., Li, W., & Huang, Y. (2021). Large group activity security risk assessment and risk early warning based on random forest algorithm. *Pattern Recognition Letters*, *144*, 1-5.
[24]. Čík, I., Rasamoelina, A. D., Mach, M., & Sinčák, P. (2021). Explaining deep neural network using layer-wise relevance propagation and integrated gradients. 2021 IEEE 19th world symposium on applied machine intelligence and informatics (SAMI),
[25]. Coulter, R., Han, Q.-L., Pan, L., Zhang, J., & Xiang, Y. (2020). Code analysis for intelligent cyber systems: A data-driven approach. *Information sciences*, *524*, 46-58.
[26]. Crawford, L., Morgan, G. G., & Cordery, C. J. (2018). Accountability and not-for-profit organisations: Implications for developing international financial reporting standards. *Financial accountability & management*, *34*(2), 181-205.
[27]. Dang, Q.-V. (2021). Improving the performance of the intrusion detection systems by the machine learning explainability. *International Journal of Web Information Systems*, *17*(5), 537-555.
[28]. Das, S., & Shiva, S. (2021). Machine learning application lifecycle augmented with explanation and security. 2021 IEEE 12th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON),
[29]. Das, T., Shukla, R. M., & Sengupta, S. (2021). The devil is in the details: Confident & explainable anomaly detector for software-defined networks. 2021 IEEE 20th International Symposium on Network Computing and Applications (NCA),
[30]. Dias, T., Oliveira, N., Sousa, N., Praça, I., & Sousa, O. (2021). A hybrid approach for an interpretable and explainable intrusion detection system. International Conference on Intelligent Systems Design and Applications,
[31]. Durocher, S., Fortin, A., Allini, A., & Zagaria, C. (2019). Users' legitimacy perceptions about standard-setting processes. *Accounting and Business Research*, *49*(2), 206-243.

[32]. El-Sappagh, S., Alonso, J. M., Islam, S. R., Sultan, A. M., & Kwak, K. S. (2021). A multilayer multimodal detection and prediction model based on explainable artificial intelligence for Alzheimer's disease. *Scientific reports*, *11*(1), 2660.
[33]. Elshawi, R., Al-Mallah, M. H., & Sakr, S. (2019). On the interpretability of machine learning-based model for predicting hypertension. *BMC medical informatics and decision making*, *19*(1), 146.
[34]. Fan, C., Xiao, F., Yan, C., Liu, C., Li, Z., & Wang, J. (2019). A novel methodology to explain and evaluate data-driven building energy performance models based on interpretable machine learning. *Applied Energy*, *235*, 1551-1560.
[35]. Geluvaraj, B., Satwik, P., & Ashok Kumar, T. (2018). The future of cybersecurity: Major role of artificial intelligence, machine learning, and deep learning in cyberspace. International Conference on Computer Networks and Communication Technologies: ICCNCT 2018,
[36]. Gorzałczany, M. B., Piekoszewski, J., & Rudziński, F. (2020). A modern data-mining approach based on genetically optimized fuzzy systems for interpretable and accurate smart-grid stability prediction. *Energies*, *13*(10), 2559.
[37]. Guerra-Manzanares, A., Nõmm, S., & Bahsi, H. (2019). Towards the integration of a post-hoc interpretation step into the machine learning workflow for IoT botnet detection. 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA),
[38]. Habibullah, S. M., & Md. Foysal, H. (2021). A Data Driven Cyber Physical Framework For Real Time Production Control Integrating IOT And Lean Principles. *American Journal of Interdisciplinary Studies*, *2*(03), 35–70. https://doi.org/10.63125/20nhqs87
[39]. Hariharan, S., Velicheti, A., Anagha, A., Thomas, C., & Balakrishnan, N. (2021). Explainable artificial intelligence in cybersecurity: A brief review. 2021 4th International Conference on Security and Privacy (ISEA-ISAP),
[40]. Henderson, D., & O'Brien, P. C. (2017). The standard-setters' toolkit: can principles prevail over bright lines? *Review of Accounting Studies*, *22*(2), 644-676.
[41]. Hernandes, P. R. G., Floret, C. P., De Almeida, K. F. C., Da Silva, V. C., Papa, J. P., & Da Costa, K. A. P. (2021). Phishing detection using URL-based XAI techniques. 2021 IEEE Symposium Series on Computational Intelligence (SSCI),
[42]. Hernandez, C. S., Ayo, S., & Panagiotakopoulos, D. (2021). An explainable artificial intelligence (xAI) framework for improving trust in automated ATM tools. 2021 IEEE/AIAA 40th Digital Avionics Systems Conference (DASC),
[43]. Himick, D., & Brivot, M. (2018). Carriers of ideas in accounting standard-setting and financialization: The role of epistemic communities. *Accounting, Organizations and Society*, *66*, 29-44.
[44]. Holder, E., & Wang, N. (2021). Explainable artificial intelligence (XAI) interactively working with humans as a junior cyber analyst. *Human-Intelligent Systems Integration*, *3*(2), 139-153.
[45]. Hsieh, C.-H., Chao, W.-C., Liu, P.-W., & Li, C.-W. (2015). Cyber security risk assessment using an interpretable evolutionary fuzzy scoring system. 2015 International Carnahan Conference on Security Technology (ICCST),
[46]. Hu, H., Liu, Y., Chen, C., Zhang, H., & Liu, Y. (2020). Optimal decision making approach for cyber security defense using evolutionary game. *IEEE Transactions on Network and Service Management*, *17*(3), 1683-1700.
[47]. Hussinger, K., & Schwiebacher, F. (2015). The market value of technology disclosures to standard setting organizations. *Industry and Innovation*, *22*(4), 321-344.
[48]. Iadarola, G., Martinelli, F., Mercaldo, F., & Santone, A. (2021). Towards an interpretable deep learning model for mobile malware detection and family identification. *Computers & Security*, *105*, 102198.
[49]. Jaeger, S. R., Spinelli, S., Ares, G., & Monteleone, E. (2018). Linking product-elicited emotional associations and sensory perceptions through a circumplex model based on valence and arousal: Five consumer studies. *Food research international*, *109*, 626-640.
[50]. Jaigirdar, F. T., Rudolph, C., Oliver, G., Watts, D., & Bain, C. (2020). What information is required for explainable AI?: A provenance-based research agenda and future challenges. 2020 IEEE 6th International Conference on Collaboration and Internet Computing (CIC),
[51]. Jiménez-Luna, J., Grisoni, F., & Schneider, G. (2020). Drug discovery with explainable artificial intelligence. *Nature Machine Intelligence*, *2*(10), 573-584.
[52]. Jones, E., & Knaack, P. (2019). Global financial regulation: Shortcomings and reform options. *Global Policy*, *10*(2), 193-206.
[53]. Kamath, U., & Liu, J. (2021). Introduction to interpretability and explainability. In *Explainable artificial intelligence: An introduction to interpretable machine learning* (pp. 1-26). Springer.
[54]. Khan, I. A., Moustafa, N., Pi, D., Sallam, K. M., Zomaya, A. Y., & Li, B. (2021). A new explainable deep learning framework for cyber threat discovery in industrial IoT networks. *IEEE Internet of Things Journal*, *9*(13), 11604-11613.
[55]. Kharlamova, N., Hashemi, S., & Træholt, C. (2021). Data-driven approaches for cyber defense of battery energy storage systems. *Energy and AI*, *5*, 100095.
[56]. Kidwell, L., & Lowensohn, S. (2018). Stakeholder participation in the governmental accounting standard-setting process. *Journal of Public Budgeting, Accounting & Financial Management*, *30*(2), 252-268.
[57]. Kim, T., Sharda, S., Zhou, X., & Pendyala, R. M. (2020). A stepwise interpretable machine learning framework using linear regression (LR) and long short-term memory (LSTM): City-wide demand-side prediction of yellow taxi and for-hire vehicle (FHV) service. *Transportation Research Part C: Emerging Technologies*, *120*, 102786.
[58]. Kuppa, A., & Le-Khac, N.-A. (2020). Black box attacks on explainable artificial intelligence (XAI) methods in cyber security. 2020 International Joint Conference on neural networks (IJCNN),
[59]. Kute, D. V., Pradhan, B., Shukla, N., & Alamri, A. (2021). Deep learning and explainable artificial intelligence techniques applied for detecting money laundering–a critical review. *IEEE access*, *9*, 82300-82317.

[60]. Li, J.-h. (2018). Cyber security meets artificial intelligence: a survey. *Frontiers of Information Technology & Electronic Engineering*, *19*(12), 1462-1474.
[61]. Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2020). Explainable ai: A review of machine learning interpretability methods. *Entropy*, *23*(1), 18.
[62]. Liu, H., Lang, B., Chen, S., & Yuan, M. (2021). Interpretable deep learning method for attack detection based on spatial domain attention. 2021 IEEE Symposium on Computers and Communications (ISCC),
[63]. Liu, H., Zhong, C., Alnusair, A., & Islam, S. R. (2021). FAIXID: A framework for enhancing AI explainability of intrusion detection results using data cleaning techniques. *Journal of network and systems management*, *29*(4), 40.
[64]. Lopez-Restrepo, S., Garcia-Tirado, J., & Alvarez, H. (2020). A MEthOdOlOGy FOR IdENtiFyiNG PhENOMENOlOGicAl-BASEd MOdElS USiNG A PARAMEtER HiERARchy. *The Canadian Journal of Chemical Engineering*, *98*(1), 213-224.
[65]. Lötsch, J., Kringel, D., & Ultsch, A. (2021). Explainable artificial intelligence (XAI) in biomedicine: Making AI decisions trustworthy for physicians and patients. *BioMedInformatics*, *2*(1), 1-17.
[66]. Loyola-Gonzalez, O., Gutierrez-Rodríguez, A. E., Medina-Pérez, M. A., Monroy, R., Martínez-Trinidad, J. F., Carrasco-Ochoa, J. A., & Garcia-Borroto, M. (2020). An explainable artificial intelligence model for clustering numerical databases. *IEEE access*, *8*, 52370-52384.
[67]. Lu, Y., Garcia, R., Hansen, B., Gleicher, M., & Maciejewski, R. (2017). The state-of-the-art in predictive visual analytics. Computer Graphics Forum,
[68]. Lyu, D., Yang, F., Kwon, H., Dong, W., Yilmaz, L., & Liu, B. (2021). Tdm: trustworthy decision-making via interpretability enhancement. *IEEE Transactions on Emerging Topics in Computational Intelligence*, *6*(3), 450-461.
[69]. Mahbooba, B., Timilsina, M., Sahal, R., & Serrano, M. (2021). Explainable artificial intelligence (XAI) to enhance trust management in intrusion detection systems using decision tree model. *Complexity*, *2021*(1), 6634811.
[70]. Maniruzzaman, B., Mohammad Anisur, R., Afrin Binta, H., Md, A., & Anisur, R. (2023). Advanced Analytics And Machine Learning For Revenue Optimization In The Hospitality Industry: A Comprehensive Review Of Frameworks. *American Journal of Scholarly Research and Innovation*, *2*(02), 52-74. https://doi.org/10.63125/8xbkma40
[71]. Mashrur, A., Luo, W., Zaidi, N. A., & Robles-Kelly, A. (2020). Machine learning for financial risk management: a survey. *IEEE access*, *8*, 203203-203223.
[72]. Mathews, S. M. (2019). Explainable artificial intelligence applications in NLP, biomedical, and malware classification: a literature review. Intelligent computing-proceedings of the computing conference,
[73]. McConville, D., & Cordery, C. (2018). Charity performance reporting, regulatory approaches and standard-setting. *Journal of Accounting and Public Policy*, *37*(4), 300-314.
[74]. Md Al Amin, K. (2022). Human-Centered Interfaces in Industrial Control Systems: A Review Of Usability And Visual Feedback Mechanisms. *Review of Applied Science and Technology*, *1*(04), 66-97. https://doi.org/10.63125/gr54qy93
[75]. Md Arif Uz, Z., & Elmoon, A. (2023). Adaptive Learning Systems For English Literature Classrooms: A Review Of AI-Integrated Education Platforms. *International Journal of Scientific Interdisciplinary Research*, *4*(3), 56-86. https://doi.org/10.63125/a30ehr12
[76]. Md Ariful, I. (2022). Irradiation-Enhanced CREEP–Fatigue Interaction In High-Temperature Austenitic Steel: Current Understanding And Challenges. *American Journal of Advanced Technology and Engineering Solutions*, *2*(04), 148-181. https://doi.org/10.63125/e46gja61
[77]. Md Nahid, H. (2022). Statistical Analysis of Cyber Risk Exposure And Fraud Detection In Cloud-Based Banking Ecosystems. *ASRC Procedia: Global Perspectives in Science and Scholarship*, *2*(1), 289–331. https://doi.org/10.63125/9wf91068
[78]. Md Sarwar, H. (2021). Sustainable Materials Characterization For Low-Carbon Construction And Infrastructure Durability. *American Journal of Interdisciplinary Studies*, *2*(01), 01-34. https://doi.org/10.63125/wq1wdr64
[79]. Md Sarwar Hossain, S., & Md Milon, M. (2022). Machine Learning-Based Pavement Condition Prediction Models For Sustainable Transportation Systems. *American Journal of Interdisciplinary Studies*, *3*(01), 31–64. https://doi.org/10.63125/1jsmkg92
[80]. Md. Mominul, H., Masud, R., & Md. Milon, M. (2022). Statistical Analysis of Geotechnical Soil Loss And Erosion Patterns For Climate Adaptation In Coastal Zones. *American Journal of Interdisciplinary Studies*, *3*(03), 36-67. https://doi.org/10.63125/xytn3e23
[81]. Md. Musfiqur, R., & Saba, A. (2021). Data-Driven Decision Support in Information Systems: Strategic Applications In Enterprises. *International Journal of Scientific Interdisciplinary Research*, *2*(2), 01-33. https://doi.org/10.63125/cfvg2v45
[82]. Md. Rabiul, K., & Sai Praveen, K. (2022). The Influence of Statistical Models For Fraud Detection In Procurement And International Trade Systems. *American Journal of Interdisciplinary Studies*, *3*(04), 203-234. https://doi.org/10.63125/9htnv106
[83]. Md. Redwanul, I., Md Nahid, H., & Md. Zahid Hasan, T. (2021). Predictive Analytics in Supply Chain Management A Review Of Business Analyst-Led Optimization Tools. *Review of Applied Science and Technology*, *6*(1), 34-73. https://doi.org/10.63125/5aypx555
[84]. Md. Tarek, H. (2023). Quantitative Risk Modeling For Data Loss And Ransomware Mitigation In Global Healthcare And Pharmaceutical Systems. *International Journal of Scientific Interdisciplinary Research*, *4*(3), 87-116. https://doi.org/10.63125/8wk2ch14

[85]. Md. Tarek, H., & Sai Praveen, K. (2021). Data Privacy-Aware Machine Learning and Federated Learning: A Framework For Data Security. *American Journal of Interdisciplinary Studies*, *2*(03), 01–34. https://doi.org/10.63125/vj1hem03

[86]. Mehdiyev, N., Houy, C., Gutermuth, O., Mayer, L., & Fettke, P. (2021). Explainable artificial intelligence (XAI) supporting public administration processes–on the potential of XAI in tax audit processes. International Conference on Wirtschaftsinformatik,

[87]. Mhlanga, D. (2021). Financial inclusion in emerging economies: The application of machine learning and artificial intelligence in credit risk assessment. *International journal of financial studies*, *9*(3), 39.

[88]. Mohammad Mushfequr, R., & Ashraful, I. (2023). Automation And Risk Mitigation in Healthcare Claims: Policy And Compliance Implications. *Review of Applied Science and Technology*, *2*(04), 124–157. https://doi.org/10.63125/v73gyg14

[89]. Molnar, C., Casalicchio, G., & Bischl, B. (2020). Interpretable machine learning–a brief history, state-of-the-art and challenges. Joint European conference on machine learning and knowledge discovery in databases,

[90]. Mst. Shahrin, S., & Samia, A. (2023). High-Performance Computing For Scaling Large-Scale Language And Data Models In Enterprise Applications. *ASRC Procedia: Global Perspectives in Science and Scholarship*, *3*(1), 94–131. https://doi.org/10.63125/e7yfwm87

[91]. Murino, G., Armando, A., & Tacchella, A. (2019). Resilience of cyber-physical systems: an experimental appraisal of quantitative measures. 2019 11th international conference on cyber conflict (CyCon),

[92]. Nascita, A., Montieri, A., Aceto, G., Ciuonzo, D., Persico, V., & Pescapé, A. (2021). XAI meets mobile traffic classification: Understanding and improving multimodal deep learning architectures. *IEEE Transactions on Network and Service Management*, *18*(4), 4225-4246.

[93]. Nguyen, T. T., & Reddi, V. J. (2021). Deep reinforcement learning for cyber security. *IEEE Transactions on Neural Networks and Learning Systems*, *34*(8), 3779-3795.

[94]. Nicholls, J., Kuppa, A., & Le-Khac, N.-A. (2021). Financial cybercrime: A comprehensive survey of deep learning approaches to tackle the evolving financial crime landscape. *IEEE access*, *9*, 163965-163986.

[95]. Nikou, M., Mansourfar, G., & Bagherzadeh, J. (2019). Stock price prediction using DEEP learning algorithm and its comparison with machine learning algorithms. *Intelligent Systems in Accounting, Finance and Management*, *26*(4), 164-174.

[96]. Noor, U., Anwar, Z., Amjad, T., & Choo, K.-K. R. (2019). A machine learning-based FinTech cyber threat attribution framework using high-level indicators of compromise. *Future Generation Computer Systems*, *96*, 227-242.

[97]. Nor, A. K. M., Pedapati, S. R., Muhammad, M., & Leiva, V. (2021). Overview of explainable artificial intelligence for prognostic and health management of industrial assets based on preferred reporting items for systematic reviews and meta-analyses. *Sensors*, *21*(23), 8020.

[98]. Norel, M., Krawiec, K., & Kundzewicz, Z. W. (2021). Machine learning modeling of climate variability impact on river runoff. *Water*, *13*(9), 1177.

[99]. Oconitrillo, L. R. R., Vargas, J. J., Camacho, A., Burgos, A., & Corchado, J. M. (2021). RYEL System: A novel method for capturing and represent knowledge in a legal domain using explainable artificial intelligence (XAI) and granular computing (GrC). In *Interpretable Artificial Intelligence: A Perspective of Granular Computing* (pp. 369-399). Springer.

[100]. Omar Muhammad, F., & Md Redwanul, I. (2023). A Quantitative Study on AI-Driven Employee Performance Analytics In Multinational Organizations. *American Journal of Interdisciplinary Studies*, *4*(04), 145-176. https://doi.org/10.63125/vrsjp515

[101]. Omar Muhammad, F., & Md. Redwanul, I. (2023). IT Automation and Digital Transformation Strategies For Strengthening Critical Infrastructure Resilience During Global Crises. *American Journal of Interdisciplinary Studies*, *4*(04), 145-176. https://doi.org/10.63125/vrsjp515

[102]. Omar Muhammad, F., & Mst. Shahrin, S. (2021). Comparative Analysis of BI Systems In The U.S. And Europe: Lessons In Data Governance And Predictive Analytics. *Journal of Sustainable Development and Policy*, *1*(5), 01-38. https://doi.org/10.63125/6b3aeg93

[103]. Palatnik de Sousa, I., Vellasco, M. M., & Costa da Silva, E. (2021). Explainable artificial intelligence for bias detection in covid ct-scan classifiers. *Sensors*, *21*(16), 5657.

[104]. Peng, J., Zou, K., Zhou, M., Teng, Y., Zhu, X., Zhang, F., & Xu, J. (2021). An explainable artificial intelligence framework for the deterioration risk prediction of hepatitis patients. *Journal of medical systems*, *45*(5), 61.

[105]. Pereira, A., & Thomas, C. (2020). Challenges of machine learning applied to safety-critical cyber-physical systems. *Machine Learning and Knowledge Extraction*, *2*(4), 579-602.

[106]. Pérez-Landa, G. I., Loyola-González, O., & Medina-Pérez, M. A. (2021). An explainable artificial intelligence model for detecting xenophobic tweets. *Applied Sciences*, *11*(22), 10801.

[107]. Pramod, A., Naicker, H. S., & Tyagi, A. K. (2021). Machine learning and deep learning: Open issues and future research directions for the next 10 years. *Computational analysis and deep learning for medical care: Principles, methods, and applications*, 463-490.

[108]. Puri, A., & Ray, S. (2020). Interpretable Machine Learning Using Switched Linear Models for Security of Cyber-Physical Systems. 2020 Integrated Communications Navigation and Surveillance Conference (ICNS),

[109]. Rakibul, H., & Samia, A. (2022). Information System-Based Decision Support Tools: A Systematic Review Of Strategic Applications In Service-Oriented Enterprises. *Review of Applied Science and Technology*, *1*(04), 26-65. https://doi.org/10.63125/w3cevz78

[110]. Rathore, H., Agarwal, S., Sahay, S. K., & Sewak, M. (2018). Malware detection using machine learning and deep learning. International Conference on Big Data Analytics,
[111]. Ratul, Q. E. A., Serra, E., & Cuzzocrea, A. (2021). Evaluating attribution methods in machine learning interpretability. 2021 IEEE International Conference on Big Data (Big Data),
[112]. Razia, S. (2023). AI-Powered BI Dashboards In Operations: A Comparative Analysis For Real-Time Decision Support. *ASRC Procedia: Global Perspectives in Science and Scholarship*, *3*(1), 62–93. https://doi.org/10.63125/wqd2t159
[113]. Roessner, V., Rothe, J., Kohls, G., Schomerus, G., Ehrlich, S., & Beste, C. (2021). Taming the chaos?! Using eXplainable Artificial Intelligence (XAI) to tackle the complexity in mental health research. *European Child & Adolescent Psychiatry*, *30*(8), 1143-1146.
[114]. Roscher, R., Bohn, B., Duarte, M. F., & Garcke, J. (2020). Explainable machine learning for scientific insights and discoveries. *IEEE access*, *8*, 42200-42216.
[115]. Ryo, M., Angelov, B., Mammola, S., Kass, J. M., Benito, B. M., & Hartig, F. (2021). Explainable artificial intelligence enhances the ecological interpretability of black-box species distribution models. *Ecography*, *44*(2), 199-205.
[116]. Sahakyan, M., Aung, Z., & Rahwan, T. (2021). Explainable artificial intelligence for tabular data: A survey. *IEEE access*, *9*, 135392-135422.
[117]. Sahlaoui, H., Nayyar, A., Agoujil, S., & Jaber, M. M. (2021). Predicting and interpreting student performance using ensemble models and shapley additive explanations. *IEEE access*, *9*, 152688-152703.
[118]. Saikat, S. (2021). Real-Time Fault Detection in Industrial Assets Using Advanced Vibration Dynamics And Stress Analysis Modeling. *American Journal of Interdisciplinary Studies*, *2*(04), 39–68. https://doi.org/10.63125/0h163429
[119]. Saikat, S. (2022). CFD-Based Investigation of Heat Transfer Efficiency In Renewable Energy Systems. *International Journal of Scientific Interdisciplinary Research*, *1*(01), 129–162. https://doi.org/10.63125/ttw40456
[120]. Samek, W., Montavon, G., Lapuschkin, S., Anders, C. J., & Müller, K.-R. (2021). Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE*, *109*(3), 247-278.
[121]. Sarker, I. H. (2021). Deep cybersecurity: a comprehensive overview from neural network and deep learning perspective. *SN Computer Science*, *2*(3), 154.
[122]. Serban, A., van der Blom, K., Hoos, H., & Visser, J. (2021). Practices for engineering trustworthy machine learning applications. 2021 IEEE/ACM 1st Workshop on AI engineering-software engineering for AI (WAIN),
[123]. Shaikh, S., & Aditya, D. (2021). Federated Learning-Driven Predictive Quality Analytics and Supply Chain Optimization In Distributed Manufacturing Networks. *Review of Applied Science and Technology*, *6*(1), 74-107. https://doi.org/10.63125/k18cbz55
[124]. Sharma, N., Sharma, R., & Jindal, N. (2021). Machine learning and deep learning applications-a vision. *Global Transitions Proceedings*, *2*(1), 24-28.
[125]. Shickel, B., Loftus, T. J., Adhikari, L., Ozrazgat-Baslanti, T., Bihorac, A., & Rashidi, P. (2019). DeepSOFA: a continuous acuity score for critically ill patients using clinically interpretable deep learning. *Scientific reports*, *9*(1), 1879.
[126]. Singer, G., & Cohen, I. (2020). An objective-based entropy approach for interpretable decision tree models in support of human resource management: The case of absenteeism at work. *Entropy*, *22*(8), 821.
[127]. Singh, A., & Akhilesh, K. (2019). The insurance industry—cyber security in the hyper-connected age. In *Smart technologies: Scope and applications* (pp. 201-219). Springer.
[128]. Sokol, K., & Flach, P. (2020). One explanation does not fit all: The promise of interactive explanations for machine learning transparency. *KI-Künstliche Intelligenz*, *34*(2), 235-250.
[129]. Stadler, C., & Nobes, C. W. (2018). Accounting for government grants: Standard-setting and accounting choice. *Journal of Accounting and Public Policy*, *37*(2), 113-129.
[130]. Stepin, I., Alonso, J. M., Catala, A., & Pereira-Fariña, M. (2021). A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. *IEEE access*, *9*, 11974-12001.
[131]. Stiglic, G., Kocbek, P., Fijacko, N., Zitnik, M., Verbert, K., & Cilar, L. (2020). Interpretability of machine learning-based prediction models in healthcare. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *10*(5), e1379.
[132]. Sun, M., Konstantelos, I., & Strbac, G. (2018). A deep learning-based feature extraction framework for system security assessment. *IEEE transactions on smart grid*, *10*(5), 5007-5020.
[133]. Taylor, J. E. T., & Taylor, G. W. (2021). Artificial cognition: How experimental psychology can help generate explainable artificial intelligence. *Psychonomic bulletin & review*, *28*(2), 454-475.
[134]. Tonoy Kanti, C., & Shaikat, B. (2022). Graph Neural Networks (GNNS) For Modeling Cyber Attack Patterns And Predicting System Vulnerabilities In Critical Infrastructure. *American Journal of Interdisciplinary Studies*, *3*(04), 157-202. https://doi.org/10.63125/1ykzx350
[135]. Unceta, I., Nin, J., & Pujol, O. (2021). Differential replication for credit scoring in regulated environments. *Entropy*, *23*(4), 407.
[136]. Uyheng, J., & Carley, K. M. (2021). An identity-based framework for generalizable hate speech detection. International conference on social computing, behavioral-cultural modeling and prediction and behavior representation in modeling and simulation,
[137]. Veitch, E., & Alsos, O. A. (2021). Human-centered explainable artificial intelligence for marine autonomous surface vehicles. *Journal of Marine Science and Engineering*, *9*(11), 1227.
[138]. Wagner, C., Strohmaier, M., Olteanu, A., Kıcıman, E., Contractor, N., & Eliassi-Rad, T. (2021). Measuring algorithmically infused societies. *Nature*, *595*(7866), 197-204.

[139]. Wang, M., Zheng, K., Yang, Y., & Wang, X. (2020). An explainable machine learning framework for intrusion detection systems. *IEEE access*, *8*, 73127-73141.

[140]. Wingard, C., Bosman, J., & Amisi, B. (2016). The legitimacy of IFRS: An assessment of the influences on the due process of standard-setting. *Meditari Accountancy Research*, *24*(1), 134-156.

[141]. Xing, W., Guo, R., Petakovic, E., & Goggins, S. (2015). Participation-based student final performance prediction model through interpretable Genetic Programming: Integrating learning analytics, educational data mining and theory. *Computers in human behavior*, *47*, 168-181.

[142]. Yuan, S., & Wu, X. (2021). Deep learning for insider threat detection: Review, challenges and opportunities. *Computers & Security*, *104*, 102221.

[143]. Zayadul, H. (2023). Development Of An AI-Integrated Predictive Modeling Framework For Performance Optimization Of Perovskite And Tandem Solar Photovoltaic Systems. *International Journal of Business and Economics Insights*, *3*(4), 01–25. https://doi.org/10.63125/8xm7wa53