American Journal of Data Science and Analytics

Volume: 6; **Issue:** 10
Pages: 70-108

Explainable Machine Learning for Real-Time Financial Fraud Detection and Risk-Based Transaction Monitoring in U.S. Banking and Digital Payment Systems

Masrul Hasan[1];

[1]. Master of Business Administration in Business Analytics; Trine University, United States; Email: masrul2121@gmail.com

Doi: 10.63125/4h9v1876
Received: *13 July 2025;* ***Revised:*** *20 August 2025;* ***Accepted:*** *14 September 2025;* ***Published:*** *10 October 2025;*

Abstract

This study addresses real-time fraud detection in U.S. banking and digital payment systems, where transaction velocity, extreme class imbalance, adaptive adversaries, and supervisory expectations for transparent, risk-based decisions limit opaque or rule-only monitoring. It develops and evaluates an explainable ensemble machine-learning framework for real-time fraud scoring and risk-based transaction monitoring that integrates transaction and behavioral features, real-time processing, model explainability, and governance. A quantitative, cross-sectional, explanatory-predictive design was adopted. A structured questionnaire was administered to purposively sampled fraud-risk, compliance, model-development, and payment-operations professionals at U.S. banks, credit unions, payment processors, and fintech firms (n = 300 valid responses), and Random Forest, XGBoost, LightGBM, and a stacked ensemble were evaluated out of time on labeled transaction data (N = 2,400,000 illustrative transactions; fraud prevalence 0.62%). Analysis combined reliability and validity testing, Pearson correlation, multiple regression, cost-sensitive and ranking-based evaluation, probability calibration, latency testing, and SHAP-based explanation. Reliability was strong (Cronbach's α = .89–.93); the four constructs explained 68.1% of the variance in perceived monitoring effectiveness (TBFQ β = .36, RTPI β = .27, MEI β = .23, RGCM β = .19, all p ≤ .001); and the calibrated ensemble reached PR-AUC = 0.901 and ROC-AUC = 0.972 out of time, versus 0.887 and 0.968 for LightGBM, with recall of 0.76 at a 1% false-positive rate, a 99th-percentile latency of 118 ms, and tier fraud rates rising from 0.08% to 31.60%. Interpretable, calibrated ensemble scoring tied to explicit risk tiers can therefore strengthen alert prioritization, reduce customer friction, and support model-risk and compliance obligations as payment rails accelerate.

Keywords

Explainable Machine Learning; Real-Time Fraud Detection; Risk-Based Transaction Monitoring; Ensemble Learning; SHAP; U.S. Digital Payments

INTRODUCTION

Financial fraud has become an increasingly complex challenge for banks, financial institutions, and digital payment providers as financial transactions have become more digital, interconnected, and immediate. Financial fraud involves deliberate deception or manipulation to obtain money, financial services, payment credentials, goods, or other economic benefits through unauthorized or misleading activity. Common forms include payment-card fraud, account takeover, synthetic-identity fraud, authorized push-payment scams, and fraudulent electronic transfers (Bolton & Hand, 2002; West & Bhattacharya, 2016). The scale of the problem is substantial. In the United States, the Federal Trade Commission (FTC) reported more than $10 billion in consumer fraud losses in 2023, while the Federal Bureau of Investigation's Internet Crime Complaint Center (FBI IC3) reported more than $12.5 billion in losses from internet-enabled crime during the same year (FTC, 2024; FBI, 2024). These figures represent reported incidents and therefore likely understate the total impact of fraud. Financial institutions also face indirect costs, including investigation, customer reimbursement, chargebacks, regulatory exposure, operational disruption, and loss of customer trust. Therefore, effective fraud monitoring should be evaluated not only by the amount of fraud detected but also by false positives, operational workload, customer impact, and overall economic cost (Dal Pozzolo et al., 2014; Ngai et al., 2011). The operating environment for financial transactions has changed significantly with the growth of card-not-present commerce, mobile wallets, peer-to-peer transfers, online banking, and instant-payment systems. Services such as The Clearing House's Real-Time Payments (RTP) network and the Federal Reserve's FedNow Service have reduced the time between transaction initiation and settlement, leaving financial institutions with limited opportunities to identify and stop fraudulent transactions. In many cases, decisions must be made within a very short processing window using only information available at authorization (Mashiat, 2024; Sumaiya, 2024; Syed Feroz, 2024). This has increased the importance of real-time monitoring systems capable of analyzing transactions as they occur. Streaming-based fraud-detection architectures have therefore become increasingly important because traditional batch processing may not satisfy real-time requirements(Carcillo et al., 2018).

The speed of these payment environments also creates a difficult balance between security and convenience. A system that blocks too many legitimate transactions can create customer frustration, while a system that allows suspicious transactions to proceed may increase financial losses. The expansion of digital payment channels has also increased the complexity of identifying fraudulent behavior. Fraudulent transactions may originate from compromised accounts, stolen credentials, manipulated devices, or legitimate accounts controlled through social-engineering techniques. As a result, fraud may not always appear as an obviously abnormal transaction. A fraudulent payment can sometimes resemble normal customer behavior, particularly when attackers have obtained sufficient information about the victim. This makes it increasingly important to analyze multiple transaction and behavioral characteristics together rather than relying only on individual rules or thresholds. At the same time, fraudsters continuously adapt their strategies in response to security controls and detection mechanisms. They may change transaction amounts, devices, locations, timing, beneficiaries, or methods of social engineering in order to avoid detection. As fraud patterns evolve, the statistical relationships learned from historical data may also change, creating the problem of concept drift (Gama et al., 2014; Lu et al., 2019). A model that performs well during development may therefore lose effectiveness after deployment, making continuous monitoring, evaluation, and adaptation important components of modern fraud management.Machine learning has become a major analytical approach for addressing these challenges because it can identify complex relationships across large volumes of transaction, account, device, and behavioral data (Cao, 2022; Financial Stability Board, 2017). Unlike simple rule-based systems, machine-learning models can consider multiple characteristics simultaneously and identify nonlinear patterns that may be difficult to define manually. Tree-based ensemble methods are particularly useful for heterogeneous financial data and have demonstrated strong performance in structured prediction problems (Breiman, 2001; Chen & Guestrin, 2016; Grinsztajn et al., 2022; Ke et al., 2017). These methods can help institutions identify combinations of characteristics associated with suspicious activity and prioritize transactions for further investigation. However, strong predictive capability does not remove the practical challenges associated with financial fraud data.One of the most important challenges is severe class imbalance because legitimate

transactions normally outnumber fraudulent transactions by a very large margin. A model may therefore achieve high overall accuracy while still failing to identify many fraudulent transactions. Fraud labels can also be delayed because suspicious activity may only be confirmed after customer complaints, payment disputes, investigator review, or subsequent investigation. Furthermore, fraud teams have limited capacity to manually investigate every alert generated by a detection system. These factors make accuracy an inadequate measure of fraud-detection performance (Hand et al., 2008; He & Garcia, 2009). Measures such as precision, recall, precision-recall relationships, ranking quality, and expected cost provide a more meaningful assessment of whether a system can identify useful fraud signals while controlling unnecessary alerts (Dal Pozzolo et al., 2014; Elkan, 2001; Saito & Rehmsmeier, 2015).

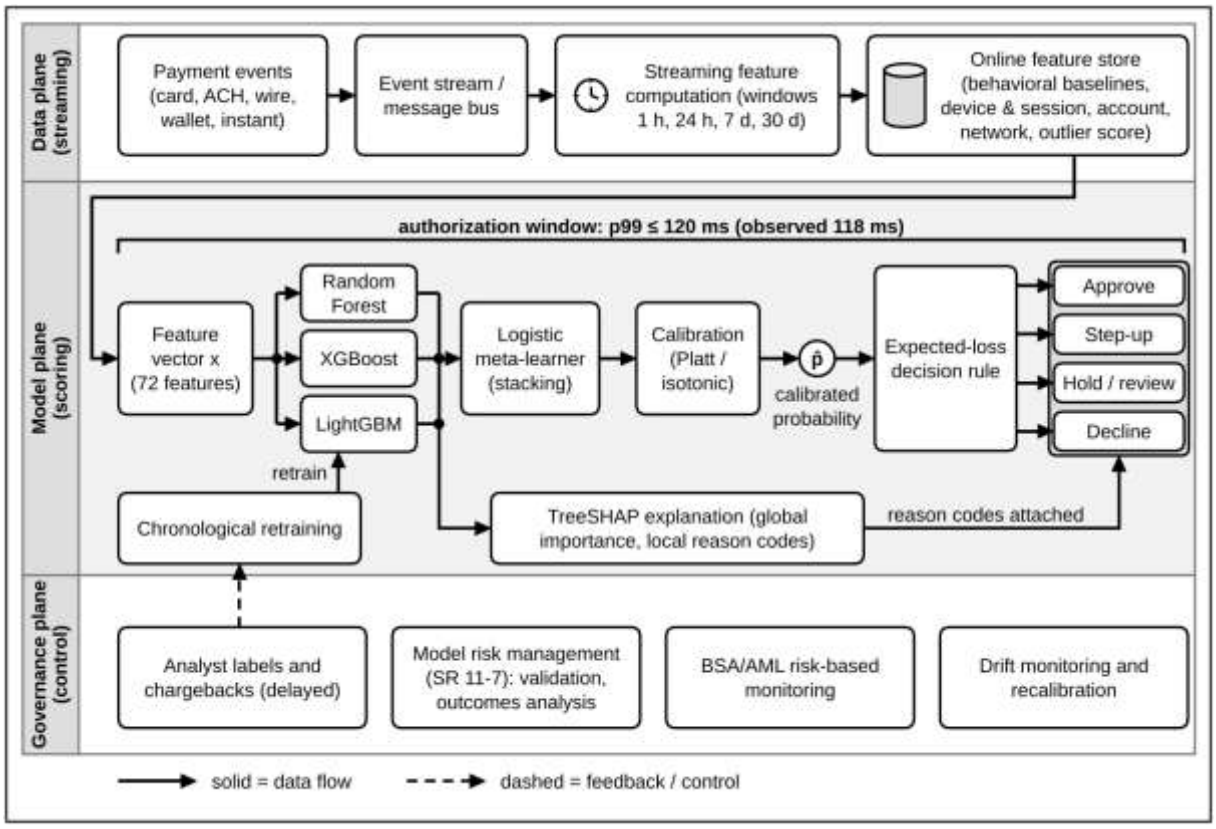
The cost of errors is also asymmetric. Missing a fraudulent transaction can result in direct financial loss, customer harm, and additional exposure to fraudulent activity. On the other hand, incorrectly blocking legitimate transactions can create customer frustration, interrupt business activity, increase investigation costs, and damage trust in the financial institution. Consequently, an effective fraud-detection system must balance fraud prevention with customer experience and operational efficiency. This makes fraud detection different from a conventional classification problem where the objective may be to maximize a single statistical performance measure. In practice, institutions need models that can prioritize high-risk transactions while maintaining manageable alert volumes and acceptable levels of customer friction. The usefulness of a model therefore depends not only on how accurately it classifies transactions but also on whether its outputs can be incorporated into real operational workflows.Another important challenge is the interpretability of complex machine-learning models. High-performing ensemble models can capture nonlinear relationships and interactions but may be difficult for fraud analysts, managers, validators, and compliance personnel to understand. This is important because fraud-detection systems operate within institutional decision processes rather than as isolated technical tools. Model outputs may influence whether transactions are approved, declined, delayed, or escalated for investigation. Financial-sector model governance has therefore emphasized the importance of understanding, validating, and appropriately challenging material models (Board of Governors of Federal Reserve System & OCC, 2011). Similarly, risk-based approaches to transaction monitoring emphasize appropriate controls, documentation, and governance (FFIEC, 2014). Explainability has consequently become an important consideration in the use of artificial intelligence in financial services. Although fraud detection is distinct from credit decision-making, regulatory developments in automated financial decisions demonstrate broader expectations for transparency and accountability. The CFPB has emphasized that the complexity of an algorithm does not remove applicable requirements to provide specific reasons for certain credit decisions (CFPB, 2022). NIST also identifies explainability and interpretability as important characteristics of trustworthy AI (NIST, 2023). These principles are relevant to fraud detection because analysts and other stakeholders may need to understand why a transaction was considered suspicious and whether the model's reasoning is consistent with the available evidence. Greater interpretability can also support communication between technical teams and decision-makers who may not have specialized knowledge of machine-learning algorithms. Explainable artificial intelligence (XAI) can help connect predictive performance with human oversight. SHAPley Additive explanations (SHAP), for example, can attribute a model prediction to individual input variables, allowing analysts to understand which factors contributed to a particular fraud-risk score (Lundberg & Lee, 2017; Lundberg et al., 2020). Such explanations can support fraud investigation, model validation, monitoring, and communication between technical and non-technical stakeholders. Previous research in financial risk has demonstrated the potential usefulness of Shapley-based explanations for risk managers and supervisors (Bracke et al., 2019; Bussmann et al., 2021). For fraud monitoring, explanations may also help investigators identify recurring characteristics of suspicious transactions and determine whether a model is responding to meaningful risk signals or unusual but legitimate behavior.

However, explainability alone does not guarantee that a model is reliable or appropriate. A model may be explainable while still suffering from poor data quality, concept drift, inappropriate assumptions, or weak governance. Similarly, an explanation of a prediction does not necessarily establish that the relationship identified by the model is causal. Therefore, explainability should be considered as one

component of a broader system of human oversight and responsible model management. Its value depends partly on whether the explanation is understandable and useful to the people who are expected to act on the model's output. This makes the connection between XAI and organizational capability particularly important in real-time fraud environments. Governance is particularly important because fraud-detection systems involve multiple organizational functions. Data scientists may develop models, information-technology teams may maintain real-time data infrastructure, fraud analysts may investigate alerts, compliance teams may assess requirements, and model-risk functions may validate performance and assumptions. Effective fraud detection therefore depends on coordination among technical and non-technical stakeholders. Governance must address issues such as model ownership, validation, documentation, performance monitoring, change management, data quality, and accountability for decisions. Federal banking agencies and FinCEN have also encouraged innovative and risk-based approaches to financial-crime monitoring while emphasizing appropriate risk management and governance (Board of Governors of Federal Reserve System et al., 2018). These governance requirements become increasingly important as institutions rely more heavily on automated systems to make decisions at high transaction volumes. Organizational capability is another important dimension of fraud detection. A technically strong model may have limited practical value if an institution lacks reliable data, real-time infrastructure, skilled personnel, investigator capacity, or effective monitoring processes. The ability to translate model outputs into useful operational decisions is therefore as important as the model itself. Human investigators also provide valuable feedback about false positives, emerging fraud patterns, and unusual legitimate behavior. This creates a feedback relationship between automated analytics and human expertise. Effective fraud management can consequently be viewed as a socio-technical capability in which machine learning, technology, human judgment, and governance operate together. The organizational environment also affects the sustainability of fraud-detection systems. Models require regular monitoring because transaction behavior, fraud strategies, customer patterns, and payment technologies can change over time. Institutions therefore need processes for identifying performance deterioration, reviewing model outputs, updating models, and documenting important changes. Data quality must also be monitored because missing, delayed, or inconsistent information can reduce model effectiveness even when the underlying algorithm remains unchanged. Similarly, investigator capacity must be considered when setting detection thresholds because an excessive number of alerts may reduce the ability of analysts to focus on genuinely high-risk cases. These considerations show that fraud detection is not simply a problem of selecting an appropriate machine-learning algorithm; it is also a problem of maintaining the infrastructure, people, processes, and controls necessary for that algorithm to create practical value. Despite the growing literature on machine-learning-based fraud detection, several research gaps remain. Many studies focus primarily on model performance using individual or public datasets without fully considering real-time processing requirements, investigator capacity, governance, and organizational conditions (Ali et al., 2022; Hilal et al., 2022). Similarly, explainability research in financial services has often focused on credit scoring rather than real-time payment fraud (Bussmann et al., 2021). Research on model governance and transparency has also frequently remained conceptually separate from the technical evaluation of fraud-detection systems (Bücker et al., 2022; Gramegna & Giudici, 2021). As a result, there is a need for an integrated perspective that considers predictive performance, explainability, governance, and organizational capability together. A further gap concerns the limited connection between technical model performance and the institutional conditions required for effective fraud monitoring. A model may perform well in an experimental environment but produce limited practical value if an organization cannot process its predictions quickly, investigate the resulting alerts, or maintain the model over time. Conversely, strong data infrastructure, skilled personnel, appropriate governance, and effective feedback mechanisms can support the practical use of analytical systems. Examining these relationships can therefore provide a more realistic understanding of fraud detection as an organizational capability rather than simply a machine-learning task. Against this background, the present study examines real-time financial fraud detection as an integrated data, machine-learning, explainability, governance, and organizational problem. The study combines a practitioner survey of individuals involved in developing or operating fraud controls with an out-of-time evaluation of an explainable ensemble model. This approach allows

institutional capabilities to be examined alongside predictive performance and interpretability. The out-of-time evaluation is particularly relevant because fraud patterns may change over time, making temporal performance an important consideration when assessing whether a model can remain useful beyond its development dataset. Rather than treating fraud detection simply as a classification problem, the study considers how technical performance interacts with the organizational conditions required to deploy, monitor, interpret, and govern fraud-detection systems effectively. The integrated approach recognizes that successful fraud monitoring requires more than a high-performing algorithm. Financial institutions must also maintain reliable data, respond to changing fraud patterns, manage alert volumes, provide meaningful explanations, support human investigation, control operational costs, and maintain appropriate governance. The study therefore contributes to a broader understanding of financial fraud detection by connecting technical modeling with practical institutional requirements. It brings together predictive performance, explainability, governance, and organizational capability within a common research direction. By doing so, it addresses the need for fraud-detection research that reflects the real operating environment of modern financial institutions, where transactions are increasingly immediate, fraud strategies are continuously changing, and automated decisions must remain understandable and subject to human oversight. Figure 1 summarizes the integrated framework developed in this study.

Figure 1: Integrated Explainable Ensemble Framework for Real-Time Fraud Detection and Risk-Based Transaction Monitoring



Background of the Study

U.S. banks, credit unions, payment processors, and fintech firms authorize payments across several channels, including card networks, automated clearing house (ACH) transfers, wire transfers, wallet-based payments, and instant-payment rails. Each channel differs in settlement speed, reversibility, liability allocation, and the data available at decision time. Fraud controls have historically combined rules, such as velocity limits and blocked-merchant lists, with scorecards and, more recently, machine-learning models. Rules are transparent and quickly modified but are brittle against new attack patterns and tend to generate many false alerts. Early systems therefore paired rules and profiling with learned models (Fawcett & Provost, 1997; Ghosh & Reilly, 1994), and later work combined classifiers trained at

different institutions and judged them with an explicit cost model (Chan et al., 1999). Machine-learning models adapt better to complex patterns but require labeled data, monitoring for drift, and clear documentation (Abdallah et al., 2016; Bolton & Hand, 2002).

Transaction-level information is rarely sufficient on its own. Aggregating a customer's recent transactions over several time windows substantially improves detection because fraud often appears as a departure from the cardholder's established behavior (Whitrow et al., 2009). Features describing the periodicity of purchase times and other behavioral regularities add further discriminating information (Bahnsen et al., 2016). Sequence models, unsupervised outlier scores, and graph-based representations of payer, payee, and merchant relationships can contribute additional signal in some settings (Carcillo et al., 2021; Jurgovsky et al., 2018; Weber et al., 2019). These data and engineering choices depend on infrastructure that can compute features in streaming mode and return a score within the authorization window, which is why real-time processing capability is treated in this study as a distinct explanatory construct. Network-based extensions that propagate suspicion through the relationships among cardholders and merchants have also improved detection (Van Vlasselaer et al., 2015), and graph neural network detectors have been designed to cope with camouflaged and imbalanced fraud structures (Dou et al., 2020; Liu et al., 2021). Explainability and governance complete the picture. Investigators need to know why a payment was flagged in order to decide quickly and consistently, customers who are declined or challenged expect a coherent process, and model validators and examiners need evidence that a model behaves sensibly and fairly (Board of Governors of the Federal Reserve System & OCC, 2011; Mehrabi et al., 2021). Explanations must be reliable as well as available, because attribution methods can be unstable or open to manipulation if used carelessly (Kumar et al., 2020; Slack et al., 2020). This study accordingly treats explainability as a measured capability and as an analytical output whose stability is itself tested. Evidence from deployed systems suggests that practitioners use explanations mostly for debugging and oversight and that different stakeholders want different things from them (Bhatt et al., 2020), and that explanation quality should be evaluated rather than assumed (Nauta et al., 2023).

Problem Statement

Fraud monitoring in U.S. banking and digital payments is frequently assessed through a single metric, such as the detection rate or the proportion of declined transactions, even though effectiveness depends on the simultaneous quality of behavioral data, speed of processing, transparency of model outputs, and maturity of governance. Conventional approaches create several difficulties. Rule-based systems cannot easily represent nonlinear interactions among transaction, device, and behavioral attributes, and they produce large volumes of false positives that burden analysts and inconvenience legitimate customers. Complex machine-learning models can capture such interactions but may operate as black boxes, which weakens analyst trust, complicates model validation, and creates difficulty in explaining outcomes to customers and supervisors (Rudin, 2019). Many published evaluations use random train-test splits and accuracy-oriented metrics that overstate performance on imbalanced and drifting data (Cerqueira et al., 2020; Kaufman et al., 2012), and they rarely report calibration, scoring latency, or the stability of explanations (Dal Pozzolo et al., 2015; Lucas & Jurgovsky, 2020). Finally, research that links institutional capabilities to monitoring effectiveness is limited, so it remains unclear which organizational and technical factors matter most. The central problem addressed by this study is the absence of an integrated, empirically tested, and interpretable framework for real-time fraud detection and risk-based transaction monitoring that connects data quality, processing infrastructure, explainability, governance, and predictive performance within the U.S. banking and digital payment context.

Purpose and Objectives of the Study

The purpose of this study is to develop and empirically evaluate an explainable ensemble machine-learning framework for real-time fraud detection and risk-based transaction monitoring in U.S. banking and digital payment systems. The study combines conventional statistical analysis of practitioner assessments with predictive modeling of transaction data so that explanatory relationships, predictive performance, and interpretability can be examined within one design. The study pursues eight objectives. The first is to determine the extent to which transaction and behavioral feature quality is associated with fraud detection and monitoring effectiveness. The second is to evaluate the relationship

between real-time processing and data-infrastructure capability and monitoring effectiveness. The third is to assess how model explainability and interpretability relate to monitoring effectiveness. The fourth is to determine the influence of risk-based governance and compliance maturity on monitoring effectiveness. The fifth is to test the combined effect of the four constructs on monitoring effectiveness using correlation and multiple regression analysis. The sixth is to build Random Forest, XGBoost, and LightGBM classifiers and a calibrated stacked ensemble from integrated transaction, behavioral, device, and network features. The seventh is to evaluate the individual and ensemble models out of time using ranking-based, cost-sensitive, calibration, and latency measures suited to imbalanced fraud data. The eighth is to apply SHAP to identify the relative importance, direction, and stability of predictors and to translate model output into risk tiers with attached reason codes.

Research Hypotheses

The hypotheses test the relationships between the principal institutional and technical constructs and Fraud Detection and Monitoring Effectiveness (FDME), using Pearson correlation and multiple regression at the .05 significance level. H1 states that Transaction and Behavioral Feature Quality (TBFQ) has a statistically significant positive relationship with FDME, reflecting the value of aggregated, behavioral, device, and network features for discriminating fraud from legitimate activity. H2 states that Real-Time Processing and Data Infrastructure (RTPI) has a statistically significant positive relationship with FDME, reflecting the need to compute features and return scores within the authorization window. H3 states that Model Explainability and Interpretability (MEI) has a statistically significant positive relationship with FDME, reflecting the role of reason codes and transparent outputs in analyst decisions and auditability. H4 states that Risk-Based Governance and Compliance Maturity (RGCM) has a statistically significant positive relationship with FDME, reflecting the contribution of model validation, tiered controls, and monitoring discipline. H5 states that TBFQ, RTPI, MEI, and RGCM jointly have a statistically significant predictive relationship with FDME. H5 therefore assesses the collective explanatory contribution of the four constructs rather than their separate effects.

LITERATURE REVIEW

The literature relevant to this study spans financial fraud typologies, statistical and machine-learning detection methods, class imbalance and concept drift, feature engineering, streaming architectures, explainable artificial intelligence, model-risk governance, and decision theory. These areas answer different questions. Fraud-detection research asks which algorithms and features separate fraudulent from legitimate transactions. Systems research asks whether those algorithms can run within the authorization window. Explainability research asks whether the outputs can be understood and trusted. Governance research asks whether the institution can validate, monitor, and defend the model. The review that follows is organized around these questions: fraud detection in the financial system, machine learning under imbalance and drift, features and real-time infrastructure, explainability, governance and fairness, the theoretical framework, and the conceptual framework.

Financial Fraud Detection and Risk-Based Monitoring

Early reviews established the core characteristics of the fraud-detection problem. Statistical fraud detection research has stressed that fraud is rare, evolves in response to detection, and requires methods that combine supervised classification with the identification of unusual behavior (Bolton & Hand, 2002). Data-mining-based fraud detection research has also highlighted the scarcity of real labeled data and the pervasive effect of class imbalance (Phua et al., 2010). A classification framework for financial fraud covering banking, insurance, securities, and related categories found that techniques such as logistic models, neural networks, and decision trees were widely applied (Ngai et al., 2011). Later reviews extended this work to newer techniques, including computational-intelligence approaches across fraud types (West & Bhattacharya, 2016), fraud-detection systems and their practical challenges (Abdallah et al., 2016), machine learning for card fraud (Lucas & Jurgovsky, 2020), anomaly-detection techniques (Hilal et al., 2022), and machine-learning approaches more broadly (Ali et al., 2022). A common conclusion across these reviews is that detection performance depends on data and deployment conditions as much as on the choice of algorithm. Surveys of anomaly detection in the financial domain, of artificial intelligence in finance more generally, and of anti-money-laundering analytics reach similar conclusions about the importance of labeled data, adaptivity, and institutional context (Ahmed et al., 2016; Cao, 2022; Chen et al., 2018). Comparative studies of classifiers on card

data (Bhattacharyya et al., 2011; Dornadula & Geetha, 2019) and early neural-network (Ghosh & Reilly, 1994) and convolutional (Fu et al., 2016) approaches show that the choice of learner interacts with the data representation and the class balance. The practitioner perspective is particularly relevant for risk-based monitoring. Operational research on card-fraud detection has shown that investigators can inspect only a small number of alerts, so the quality of the highest-ranked alerts matters more than average classification accuracy, while unbalanced data and drift must be handled explicitly (Dal Pozzolo et al., 2014). Research has also formalized the interaction between alerts and feedback, showing that the labels available for training depend on which alerts investigators examine and that feedback arrives faster than delayed customer disputes (Dal Pozzolo et al., 2018). Risk-based monitoring in the regulatory sense has a related logic. Bank Secrecy Act examination guidance expects institutions to assess their own money-laundering and fraud risk and to scale monitoring and controls accordingly (Federal Financial Institutions Examination Council [FFIEC], 2014). Both literatures converge on the idea that scarce review capacity should be directed toward the transactions with the highest expected loss, which motivates the tiered, cost-sensitive decision structure used in this study. Performance criteria that ignore the cost and timing of intervention can mislead, which is why measures tailored to card-fraud tools have been proposed (Hand et al., 2008) and why decision-support systems have been designed around investigator workflows (Carminati et al., 2015).

Machine Learning under Class Imbalance and Concept Drift

Class imbalance is the most visible technical feature of fraud data. Resampling methods such as the Synthetic Minority Over-sampling Technique (Chawla et al., 2002) and generative approaches (Fiore et al., 2019) alter the training distribution, while cost-sensitive learning incorporates unequal misclassification costs directly into model training and decision thresholds **(**Bahnsen et al., 2015**;** Elkan, 2001; Dal Pozzolo et al., 2015) showed that undersampling biases predicted probabilities and that a correction is needed when the probabilities are used for decisions. This finding matters here because the study uses scores to compute expected loss, so it requires calibrated probabilities. More generally, boosted trees often produce poorly calibrated probabilities, which can be improved by post-hoc calibration (Niculescu-Mizil & Caruana, 2005). Post-hoc calibration by Platt scaling or isotonic regression (Platt, 2000; Zadrozny & Elkan, 2002) is likewise needed for modern high-capacity learners (Guo et al., 2017), and the Brier score (Brier, 1950) is the standard proper scoring rule for judging it. Evaluation under imbalance requires metrics that reflect the minority class (He & Garcia, 2009; Krawczyk, 2016). ROC analysis remains the common basis for comparing classifiers (Fawcett, 2006), although the AUC embeds an implicit, classifier-dependent weighting of misclassification costs (Hand, 2009), and the relationship between ROC and precision-recall curves is well characterized (Davis & Goadrich, 2006). Resampling methods such as SMOTE have since been extended in many directions (Fernández et al., 2018). Precision-recall curves are more informative than receiver operating characteristic (ROC) curves when positives are rare (Saito & Rehmsmeier, 2015), and ranking measures such as precision among the highest-scored alerts reflect investigator capacity (Dal Pozzolo et al., 2014). Concept drift and delayed supervision complicate both training and evaluation. (Gama et al., 2014; Lu et al., 2019) reviewed methods for detecting and adapting to changes in the data-generating process. Context-tracking learners (Widmer & Kubat, 1996) and adaptive-windowing detectors (Bifet & Gavaldà, 2007) are foundational examples. In fraud applications, drift arises because attackers change tactics and because customer behavior shifts seasonally (Žliobaitė et al., 2016). Prior work (Dal Pozzolo et al., 2018) showed that models trained only on delayed labels perform worse than systems that also exploit recent investigator feedback. Incremental learning strategies for card-fraud streams have also been compared on large real-world data sets (Lebichot et al., 2021). A direct implication for evaluation is that random cross-validation can leak future information into training and overstate performance (Bergmeir & Benítez, 2012**;** Cerqueira et al., 2020), so temporal, out-of-time validation is the appropriate benchmark for a deployed model. This study accordingly uses chronological splits and rolling-origin validation. Tree-based ensembles are strong candidates for this setting. Bagging (Breiman, 1996a) and gradient boosting (Friedman, 2001) are the two foundations on which they are built. Random Forests average many decorrelated trees to reduce variance (Breiman, 2001), and gradient-boosting systems such as XGBoost and LightGBM add trees sequentially to correct residual errors while remaining computationally efficient (Chen & Guestrin, 2016; Ke et al., 2017). Benchmarks on tabular data have

found that tree-based methods remain highly competitive with, and often superior to, deep-learning architectures on medium-sized heterogeneous datasets (Gorishniy et al., 2021; Grinsztajn et al., 2022; Shwartz-Ziv & Armon, 2022). Ensembles of fraud classifiers have also been evaluated directly; (Randhawa et al., 2018), for example, combined boosting and majority voting for card-fraud detection. Gradient boosting with ordered target statistics, as implemented in CatBoost, offers a further option for categorical-heavy payment data (Prokhorenkova et al., 2018), and large empirical comparisons of supervised learners have repeatedly placed tree ensembles among the strongest general-purpose methods (Caruana & Niculescu-Mizil, 2006). Stacked generalization trains a meta-learner on the predictions of base learners (Breiman, 1996b; Wolpert, 1992), and the broader ensemble literature explains why combining diverse learners can improve accuracy and stability (Dietterich, 2000; Zhou, 2012). Prior work (Hajek et al., 2023) reported that an XGBoost-based framework with engineered features performed well for fraud detection in mobile payment systems. Optimized gradient-boosting variants (Taha & Malebary, 2020), cost-sensitive random-forest hybrids (Nami & Shajari, 2018), feature-engineering schemes that feed deep architectures (Zhang et al., 2021), and ensembles of deep sequential models (Forough & Momtazi, 2021) have also been reported for card fraud. These findings support the choice of Random Forest, XGBoost, and LightGBM as complementary base learners and of a stacked ensemble as the final model.

Behavioral Features, Real-Time Processing, and Streaming Infrastructure

Feature quality is a recurring determinant of fraud-detection performance. Prior work (Whitrow et al., 2009) showed that aggregating a card's transactions over time windows improves detection relative to using isolated transactions. Prior work (Bahnsen et al., 2016) demonstrated that features capturing the periodic behavior of transaction times, together with aggregation, improved results when evaluated by monetary savings rather than by accuracy alone. Prior work (Jurgovsky et al., 2018) compared sequence classifiers with aggregated-feature models and found that the benefit of sequence modeling depended on the transaction type, which suggests that feature design should be matched to the channel. Prior work (Carcillo et al., 2021) found that adding unsupervised outlier scores at several levels of granularity improved supervised card-fraud detection. Outlier-detection methods are surveyed (Chandola et al., 2009), and isolation-based detectors in particular scale well to large transaction volumes **(**Liu et al., 2008**)**. In transaction-network settings, prior work (Weber et al., 2019) explored graph convolutional networks for financial forensics and reported that simpler methods using engineered features remained competitive, which indicates that graph-derived features should be evaluated against strong tabular baselines rather than assumed to be superior. Graph-based anomaly detection more broadly is surveyed (Akoglu et al., 2015). Real-time operation adds constraints that offline experiments rarely examine. Prior work (Carcillo et al., 2018) presented a scalable streaming framework for card-fraud detection that combined stream processing with a model that is retrained as labeled feedback arrives, and they emphasized throughput, latency, and concept drift as joint design requirements. For a payment authorization, the relevant measures include the median and tail latency of feature retrieval and scoring, the freshness of streaming aggregates, and the ability to scale during peak volumes. Tail latency is a property of the whole serving path rather than of the model alone, and large-scale services manage it with redundancy and request-level techniques (Dean & Barroso, 2013); general-purpose engines such as Spark supply the stream-processing substrate on which published fraud frameworks have been built (Carcillo et al., 2018; Zaharia et al., 2016). A model that is accurate offline but cannot be served inside the authorization window has no operational value. For this reason the study measures scoring latency as a performance outcome and treats real-time processing capability as an explanatory construct in the practitioner survey.

Explainable Artificial Intelligence in Financial Decision Systems

Explainable artificial intelligence provides methods and concepts for making model behavior understandable to users, analysts, auditors, regulators, and other stakeholders. A rigorous science of interpretability has been called for, while the term encompasses several distinct objectives, including transparency, interpretability, and post-hoc explanation **(**Doshi-Velez & Kim, 2017**;** Lipton, 2018**).** Broader taxonomies of explanation methods subsequently distinguished intrinsically interpretable models from post-hoc explanations and global explanations from local explanations (Guidotti et al., 2018; Adadi & Berrada, 2018; Arrieta et al., 2020). These distinctions are particularly important in

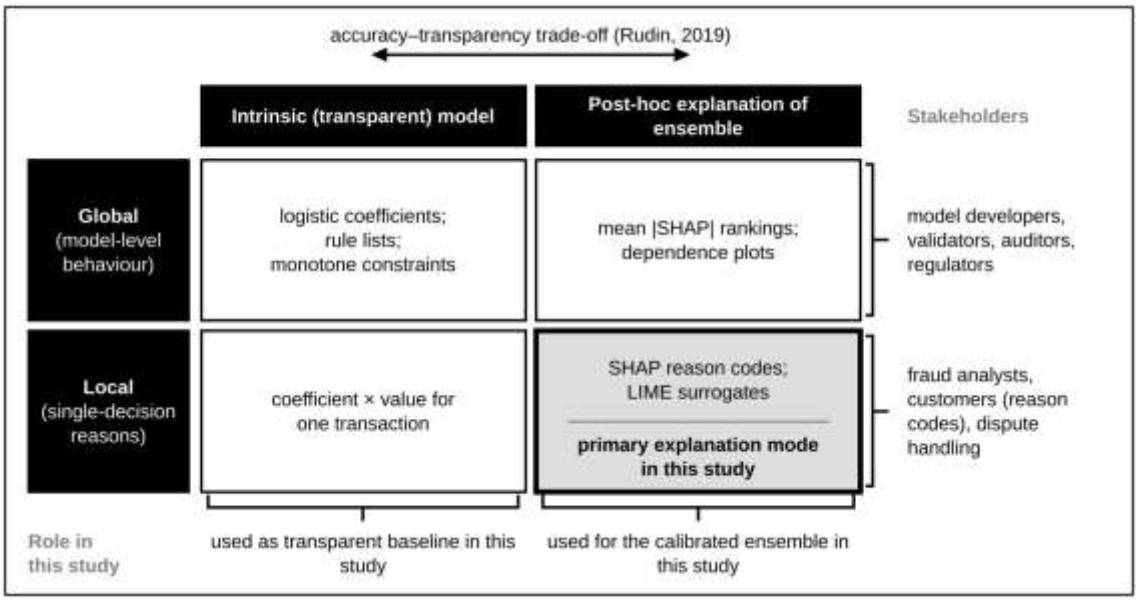
financial decision systems because different stakeholders require different levels and forms of explanation(Ashik & Shabbir, 2022; Md Imran & Jannatul, 2021). A model developer may need to understand overall feature behavior, whereas a compliance officer or financial analyst may need to understand why a particular transaction was flagged or a particular customer was assigned a higher risk score. Human explanations are generally contrastive and social, meaning that people often seek to understand why one outcome occurred rather than another (Miller, 2019). This perspective is directly relevant to financial decision-making, where analysts may need to determine why a transaction was flagged as suspicious rather than approved, or why one application received a different risk assessment from another. For high-stakes decisions, interpretable models should generally be preferred to post-hoc explanations of black-box models when they can achieve comparable predictive performance **(**Rudin, 2019). This argument provides an important methodological discipline for the present study, which therefore reports a transparent baseline alongside the ensemble model so that any improvement in predictive performance can be considered together with the associated reduction in transparency (Kazi Rakib Hasan & Adnan, 2022; Mansura Akter & Sumaiya, 2022; Mashfiqur & Md, 2022). Fundamental principles and unresolved challenges involved in developing reliable and interpretable machine-learning systems have also been identified (Rudin et al., 2022). Thus, explainability should not be treated simply as an additional visualization feature but as part of the broader process of model validation, human oversight, accountability, and responsible deployment. Among post-hoc methods, local surrogate explanations such as LIME and Shapley-value methods have been widely adopted (Ribeiro et al., 2016). SHAP unifies several attribution approaches under the Shapley value from cooperative game theory (Lundberg & Lee, 2017; Shapley, 1953) and, for tree ensembles, provides an efficient algorithm capable of producing consistent local attributions and aggregated global summaries (Lundberg et al., 2020). Shapley-style feature contributions for individual predictions had also been proposed earlier (Štrumbelj & Kononenko, 2014), while subsequent research has demonstrated that the choice of value function, including interventional versus conditional expectations, can affect the interpretation of the resulting attributions (Aas et al., 2021; Janzing et al., 2020; Sundararajan & Najmi, 2020). This is particularly relevant in financial applications because correlated variables, such as transaction amount, transaction frequency, account activity, and customer characteristics, can make attribution results difficult to interpret without considering the underlying data structure.

In credit risk, one study (Bussmann et al., 2021) applied Shapley values to explain a gradient-boosted credit model and compare risk drivers across borrowers, while other work (Bracke et al., 2019) demonstrated how Shapley-based explanations can support the review of a machine-learning default model by a central bank. Credit scoring also has one of the longest benchmarking traditions among financial prediction tasks, and a large comparison found that ensemble methods generally outperform individual classifiers (Lessmann et al., 2015). Research addressing transparency and auditability in credit scoring (Bücker et al., 2022), together with comparisons of SHAP and LIME in credit-risk applications (Gramegna & Giudici, 2021), provides a close methodological precedent for the present study. Counterfactual explanations offer a complementary approach by describing what would need to change for a different outcome to occur (Mothilal et al., 2020; Wachter et al., 2018). Such explanations can be particularly useful in financial review because they move beyond identifying influential features and instead provide information about the conditions associated with a different model decision. However, the literature also demonstrates that explanations should not automatically be treated as reliable evidence of model reasoning (Mashiat & Md Anisur Rahman, 2022; Md & Gazi Yeasin, 2022; Md Ashfaq & Abdullah Mohammad, 2022). Prior work (Slack et al., 2020) showed that LIME and SHAP explanations can be manipulated by models that behave differently on perturbed inputs, while other work (Kumar et al., 2020) identified conceptual and practical problems in interpreting Shapley values as straightforward measures of feature importance. Related studies indicate that attribution methods can be sensitive to relatively small input perturbations (Alvarez-Melis & Jaakkola, 2018**;** Ghorbani et al., 2019) and that different explanation techniques can produce substantially different interpretations of the same model (Krishna et al., 2022). These limitations motivate stability checks across time periods and cross-validation folds in the present study, as well as the treatment of explanations as supporting evidence for human review rather than proof of causation. Whether explanations actually improve human decision-making is itself an empirical question, and controlled experiments have found that the

relationship between model transparency and user performance is not necessarily straightforward (Poursabzi-Sangdeh et al., 2021). Consequently, an effective financial XAI framework should evaluate not only whether explanations can be generated, but also whether they are stable, meaningful, sufficiently understandable, and useful within the actual decision and governance process (Md Imran & Jannatul, 2022; Mst Kaniz & Jahanara, 2022; Sadia & Maishar Mostari, 2022).

Figure 2 organizes these distinctions for the fraud-monitoring setting. Intrinsically interpretable models and post-hoc explanations of complex ensembles are separated from global and local explanations, and each combination serves a different operational purpose, from model validation and audit to analyst triage and customer-facing reason codes (Adadi & Berrada, 2018; Arrieta et al., 2020; Guidotti et al., 2018; Lundberg & Lee, 2017).

Figure 2: Explainability Taxonomy Framework for Fraud Scoring in Financial Decision Systems



Governance, Compliance, and Fairness

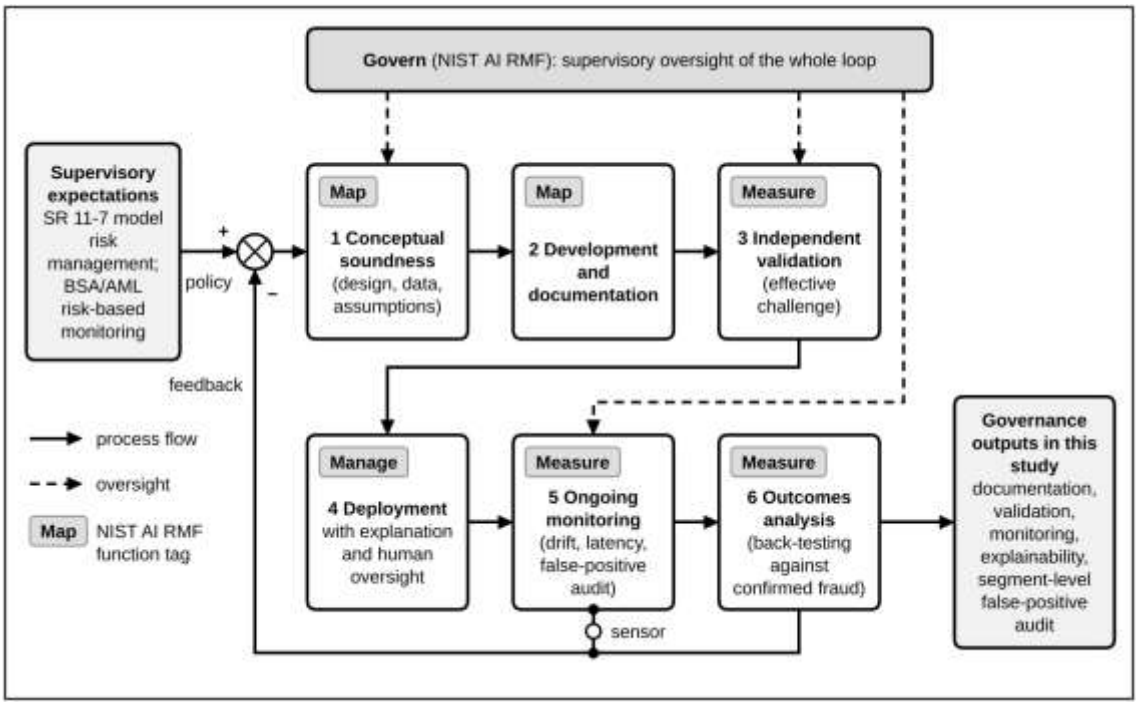
Model governance is a critical component of financial AI because predictive accuracy alone does not establish whether a model is appropriate for operational use (Shaiful et al., 2022; Stabak & Md Shahid, 2022; Syed Feroz & Md Mahamud Pervaz, 2022). Model-risk guidance issued by the Federal Reserve and the OCC defines a model broadly and requires validation covering conceptual soundness, ongoing monitoring, outcomes analysis, and effective challenge by qualified independent reviewers (Board of Governors of the Federal Reserve System & OCC, 2011). The NIST AI Risk Management Framework similarly organizes trustworthy-AI practice around the functions of govern, map, measure, and manage, while identifying accountability, transparency, explainability, and fairness as important characteristics of trustworthy systems (NIST, 2023). These principles imply that financial institutions need governance mechanisms extending across the entire model lifecycle, including data preparation, model development, validation, deployment, monitoring, modification, and eventual retirement. Financial compliance requirements provide an additional layer of control (Abdullah Mohammad, 2023; Adnan & Iftekhar, 2023; Ahmed & Nuzhat Sadia, 2023). Bank Secrecy Act examination procedures require risk-based monitoring and reporting of suspicious activity (FFIEC, 2014), meaning that automated fraud and transaction-monitoring systems must operate within established compliance processes rather than function as isolated prediction tools. In consumer credit, supervisory statements indicate that adverse-action reasons must be sufficiently specific even when complex algorithms are used (CFPB, 2022). Consequently, organizations need documentation and explanation mechanisms capable of supporting both internal review and external regulatory scrutiny. Documentation practices from the machine-learning community, such as model cards (Mitchell et al., 2019), can support this requirement by recording a model's intended use, limitations, evaluation results, and relevant performance characteristics. The OCC model risk management handbook provides additional

examination expectations concerning model development, validation, governance, and ongoing monitoring (Office of the Comptroller of the Currency [OCC], 2021). At the operational level, production machine-learning systems can also accumulate hidden maintenance costs through entanglement, undeclared dependencies, and changing data environments (Sculley et al., 2015), making continuous monitoring and controlled model change important governance responsibilities. Fairness represents a closely related governance concern because financial fraud and risk models can impose different burdens on legitimate customers when prediction errors are unevenly distributed across groups (Asheke, 2023; Gazi Yeasin, 2023; Jahanara & Mst Kaniz, 2023). A model with a low overall false-positive rate may still generate substantially higher rates of unnecessary alerts or transaction challenges for particular customer segments. Prior work (Mehrabi et al., 2021) reviewed major sources of bias and competing definitions of fairness in machine learning, while other work (Hardt et al., 2016) proposed equality of opportunity as a criterion based on comparing relevant error rates across groups. In fraud monitoring, a comparable governance question is whether false-positive rates differ systematically across customer segments, potentially resulting in disproportionate inconvenience, additional verification requirements, or delayed transactions. Because protected-class attributes are generally not collected for routine fraud screening, fairness assessment may need to rely on lawful proxies or non-protected segmentation variables, and the resulting analysis must be interpreted cautiously. Furthermore, common fairness criteria cannot always be simultaneously satisfied when groups have different base rates (Chouldechova, 2017), meaning that selecting an appropriate fairness criterion is itself a governance decision rather than a purely technical exercise. In the U.S. legal context, disparate-impact principles can also apply to algorithmic systems even where discriminatory intent is not established (Barocas & Selbst, 2016). Evidence from mortgage lending indicates that the adoption of machine-learning models can alter the distribution of benefits from credit decisions across demographic groups (Fuster et al., 2022), while other work (Kozodoi et al., 2022) demonstrate that the economic implications of fairness-aware scoring can also be examined quantitatively. These findings suggest that fairness should be incorporated into model evaluation alongside predictive performance, interpretability, and financial value.

Accordingly, the present study treats governance, compliance, and fairness as interconnected dimensions of responsible financial AI rather than as separate post-development checks (Md Anisur Rahman, 2023; Md Asif Ali Sheak & Tahmedur, 2023; Md Kaium, 2023). The proposed framework incorporates documentation, validation, monitoring, explainability, human oversight, and segment-level false-positive analysis as governance outputs. The segment-level audit is particularly important because it provides a practical mechanism for identifying whether model errors are concentrated within particular customer groups (Md Mahamud Pervaz, 2023; Mohammad Badrul & Md. Mominul, 2023; Nishat, 2023). At the same time, such an audit should be interpreted as a monitoring and risk-management instrument rather than definitive evidence of discrimination (Nuzhat Sadia, 2023; Stabak & Md Siam, 2023; Syed Feroz, 2023). By incorporating these elements into the RGCM construct, the study recognizes that a financially useful AI system must be evaluated not only according to how accurately it detects risk, but also according to whether its decisions can be explained, monitored, challenged, documented, and governed throughout its operational lifecycle.

Figure 3 summarizes the governance framework adopted in this study. Supervisory model-risk guidance emphasizes conceptual soundness, ongoing monitoring, outcomes analysis, and effective challenge (Board of Governors of the Federal Reserve System & OCC, 2011), while the NIST AI Risk Management Framework organizes trustworthy-AI practice around the govern, map, measure, and manage functions (NIST, 2023). The present study draws on both and translates them into concrete governance outputs for the fraud-monitoring model.

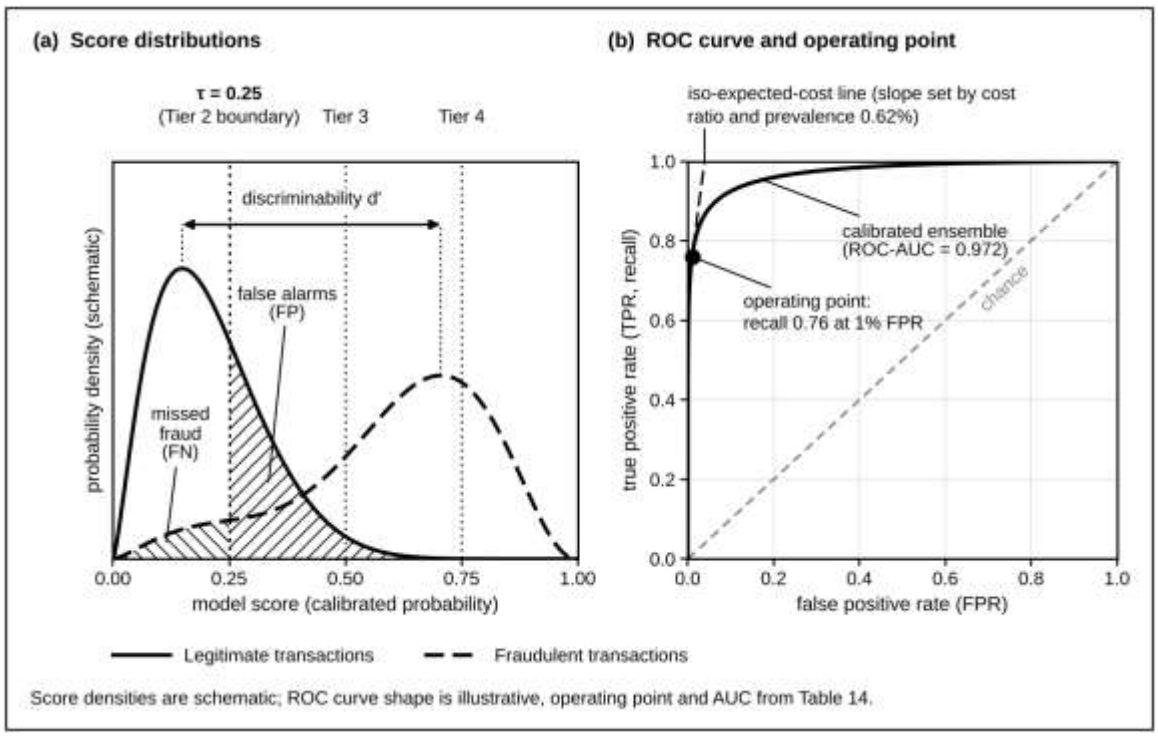
Figure 3: Model Risk Governance Framework Integrating Supervisory Guidance and Trustworthy-AI Practice



Theoretical Framework: Signal Detection and Risk-Based Decision Theory

Signal detection theory provides the primary theoretical foundation for this study. The theory describes how an observer separates signal from noise when evidence is ambiguous, with performance determined jointly by the discriminability of the evidence and by the decision criterion applied (Green & Swets, 1966). A fraud model produces a score for each transaction, the score distributions of fraudulent and legitimate transactions overlap (Figure 4), and the institution chooses a threshold that trades missed fraud against false alarms; the ROC curve is the empirical counterpart of this trade-off (Fawcett, 2006). Improving the information available to the model, through better features and timelier data, increases discriminability. Choosing the threshold in light of the relative costs of errors determines the operating point (Abdullah Mohammad, 2024; Ahmed, 2024; Tahmedur & Rifat, 2023). Cost-sensitive learning formalizes this second element: the optimal decision depends on misclassification costs as well as on probabilities (Elkan, 2001).

Figure 4: Signal Detection Framework for Fraud Scoring and Threshold Selection



In the present framework, a transaction i with amount at risk A_i and calibrated fraud probability $\hat{p_i}$ is escalated when its expected fraud loss exceeds the cost of intervention. The decision rule is:

Escalate transaction i $\Leftrightarrow \hat{p_i} \times A_i > C(review)$

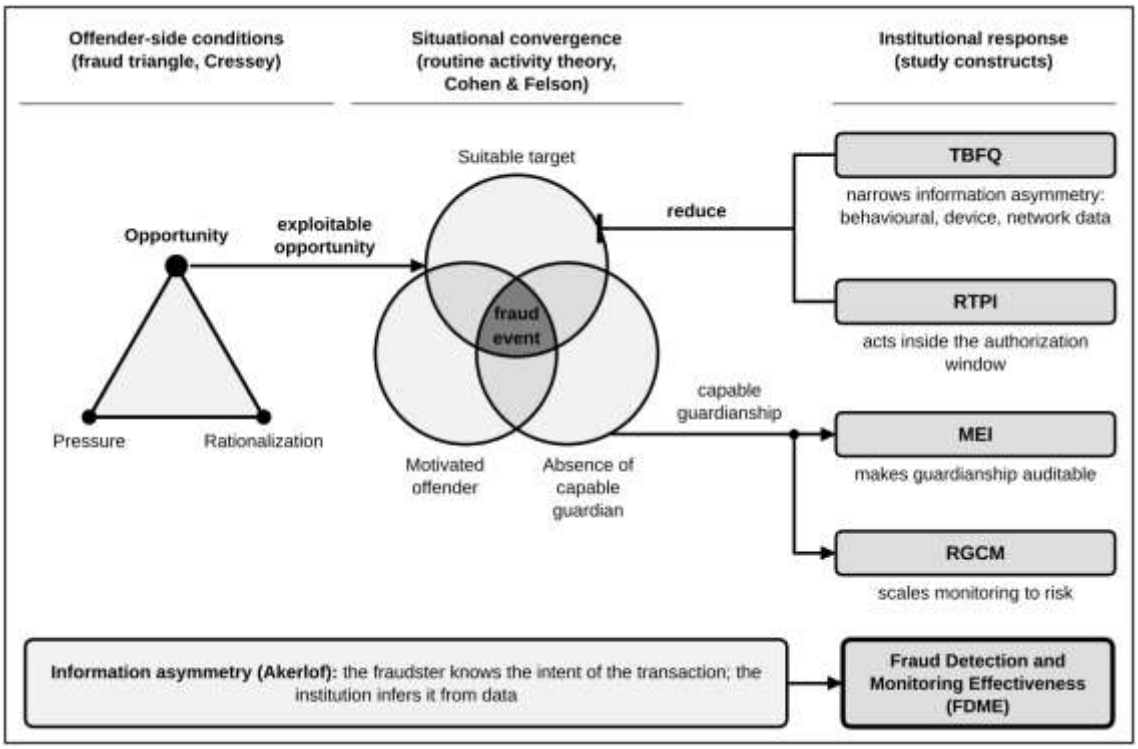
where C(review) represents the combined cost of analyst review and customer friction. Because intervention costs differ by action, the rule is applied at several thresholds, producing risk tiers that correspond to approval, step-up authentication, manual review, and decline. Expected cost across a test period is then

$E[Cost] = \Sigma_i [\, y_i \cdot (1 - d_i) \cdot A_i + d_i \cdot C(action)_i \,]$

where y_i indicates fraud, d_i indicates whether the transaction was intercepted, and $C(action)_i$ is the cost of the chosen intervention. Two complementary theories explain why the constructs in the survey should matter. Information asymmetry between institution and fraudster, in which the fraudster knows more about the true intent of a transaction than the institution, describes the problem that better behavioral and device data reduce (Akerlof, 1970). The fraud triangle of pressure, opportunity, and rationalization (Cressey, 1953) and routine activity theory, which attributes crime to the convergence of motivated offenders, suitable targets, and the absence of capable guardians (Cohen & Felson, 1979), describe fraud supply and why monitoring functions as a guardian. Finally, the technology-organization-environment framework (Tornatzky & Fleischer, 1990) supports the treatment of technical capability, organizational governance, and regulatory environment as jointly shaping the effectiveness of a deployed system (Asheke, 2024; Ashrafuzzaman, 2024; Jannatul, 2024).

Figure 5 integrates these three theoretical perspectives with the study constructs. The fraud triangle describes conditions on the offender side (Cressey, 1953), routine activity theory describes the convergence of offender, target, and guardianship (Cohen & Felson, 1979), and information asymmetry describes the knowledge gap that behavioral and device data help close (Akerlof, 1970). In this framework, TBFQ and RTPI reduce exploitable opportunity, whereas MEI and RGCM strengthen capable guardianship, and together they shape FDME.

Figure 5: Fraud Triangle, Routine Activity, and Information Asymmetry Framework Linked to the Study Constructs



The relationships tested in the survey are expressed as:

$$FDME_i = \beta_0 + \beta_1 TBFQ_i + \beta_2 RTPI_i + \beta_3 MEI_i + \beta_4 RGCM_i + \varepsilon_i$$

where TBFQ is Transaction and Behavioral Feature Quality, RTPI is Real-Time Processing and Data Infrastructure, MEI is Model Explainability and Interpretability, RGCM is Risk-Based Governance and Compliance Maturity, and ε_i is unexplained variation. The predictive component replaces the linear form with a nonlinear function learned from transaction data:

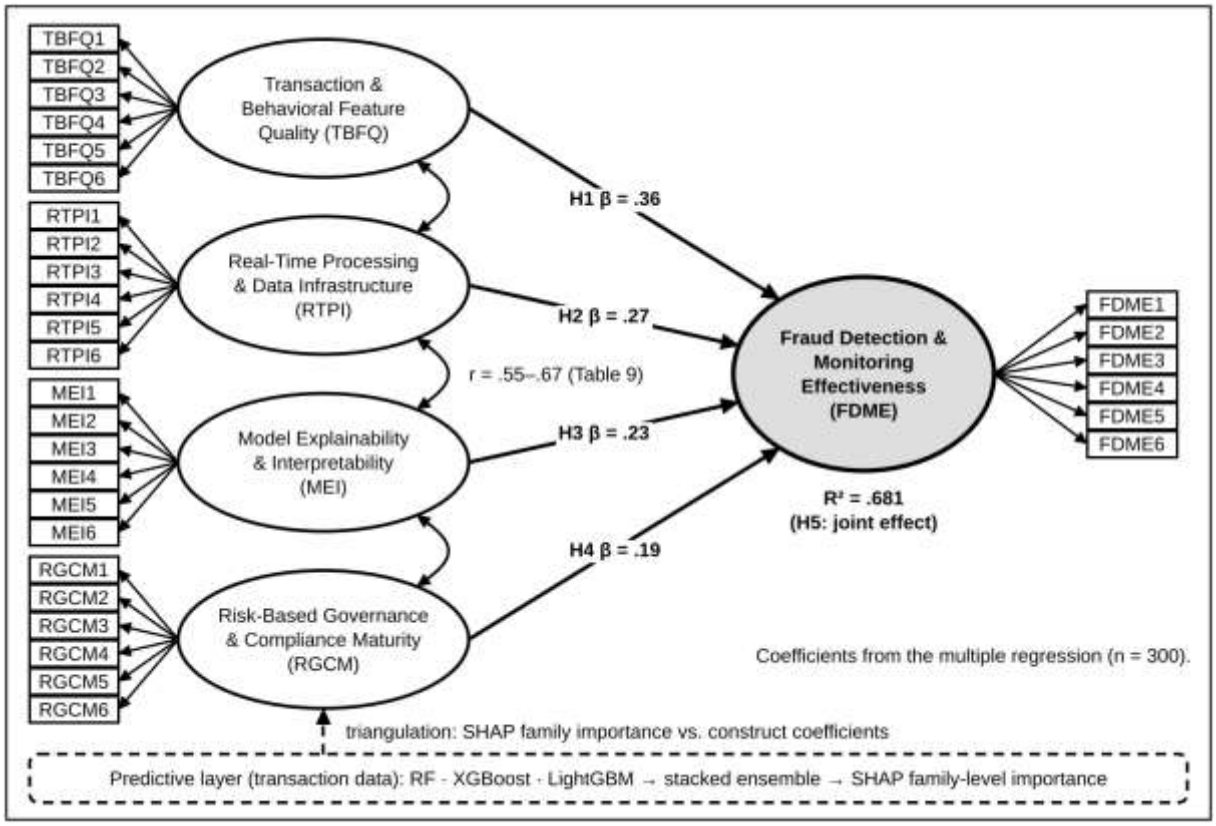
$$\hat{p_i} = F(X_i \wedge txn, X_i \wedge beh, X_i \wedge dev, X_i \wedge net)$$

where the arguments denote transaction, behavioral, device, and network feature vectors. The final ensemble combines M calibrated base learners, $\hat{p_i} \wedge ens = \Sigma_{m=1}^{M} w_m \hat{p_i}_{,m}$, $\Sigma_m w_m = 1$, $w_m \geq 0$ and the explanation of each prediction is additive: $f(x_i) = \varphi_0 + \Sigma_{j=1}^{K} \varphi_{i\,j}$ where φ_0 is the baseline output and $\varphi_{i\,j}$ is the SHAP contribution of predictor j to observation i (Lundberg & Lee, 2017). Positive SHAP values move the estimate toward fraud and negative values move it away. This structure allows regression coefficients and SHAP rankings to be interpreted within the same four-construct architecture.

Conceptual Framework for Explainable Real-Time Fraud Monitoring

The conceptual framework translates the theory into measurable constructs. The dependent construct is Fraud Detection and Monitoring Effectiveness (FDME), defined as the degree to which an institution's monitoring detects fraud promptly, limits false positives, produces high-quality alerts, and avoids unnecessary customer friction. Four explanatory constructs are specified (Figure 6). Transaction and Behavioral Feature Quality (TBFQ) captures the completeness, timeliness, and discriminating power of transaction aggregates, behavioral baselines, device and geolocation signals, and network features. Real-Time Processing and Data Infrastructure (RTPI) captures streaming data availability, feature-retrieval and scoring latency, scalability, and drift monitoring. Model Explainability and Interpretability (MEI) captures the availability of reason codes, analyst usability of explanations, auditability, and documentation of model behavior. Risk-Based Governance and Compliance Maturity (RGCM) captures model validation, tiered-control design, alignment with risk-based monitoring expectations, outcome analysis, and fairness review.

Figure 6: Conceptual Framework Linking TBFQ, RTPI, MEI, and RGCM to FDME



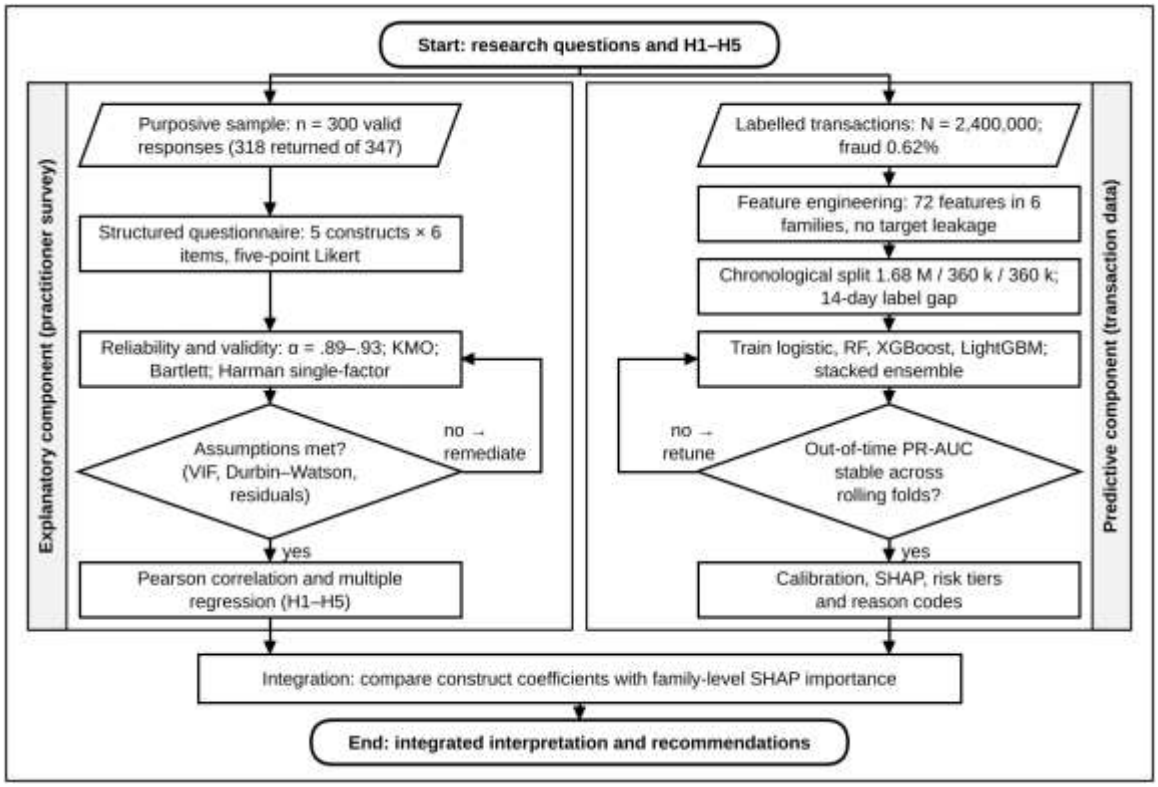
The constructs are interrelated rather than independent. Better features raise the value of faster infrastructure, explanations depend on the model and the features that feed it, and governance determines whether the other capabilities are validated and maintained. For that reason the study tests both the separate and joint contributions of the constructs. The predictive layer then examines the same domains at a finer level. Random Forest, XGBoost, and LightGBM are trained on the integrated feature set, combined into a calibrated ensemble, and explained with SHAP so that the importance of specific transaction, behavioral, device, and network variables can be compared with the construct-level regression results. The complete analytical sequence is: multimodal data and practitioner assessments, then statistical testing of H1–H5, then individual models, then calibrated ensemble prediction, then SHAP explanation, then risk-tier assignment and reason codes. Correlation and regression test H1–H5 at the construct level, and the ensemble evaluates predictive behavior across detailed variables.

METHODS

Research Design

This study adopts a quantitative, cross-sectional, explanatory-predictive design. The explanatory component uses a structured practitioner questionnaire to test H1–H5 with correlation and multiple regression. The predictive component uses labeled transaction data to train and evaluate machine-learning models out of time. The two components are linked through the four-construct architecture described in the conceptual framework, so that survey-based relationships and model-based importance can be compared. The setting is the U.S. banking and digital-payments sector, and the unit of analysis is the respondent for the survey component and the transaction (grouped by account and by time period) for the predictive component. Figure 7 summarizes the two components and how they are integrated.

Figure 7: Research Design Linking the Practitioner Survey and the Predictive Model



Population, Sampling, and Sample Size

The target population consists of professionals who design, operate, validate, or oversee fraud and transaction-monitoring controls at U.S. banks, credit unions, payment processors and networks, and fintech or wallet providers. Roles include fraud and risk analysts, data scientists and model developers, compliance and anti-money-laundering officers, payment-operations and platform engineers, and model-risk or validation staff. Purposive sampling is used because respondents require specialized knowledge (Etikan et al., 2016). A target of approximately 300 qualified responses is set. Regression guidance suggests a minimum of 50 + 8m observations for testing the overall model, where m is the number of predictors (Green, 1991), which gives 82 for four predictors, so the target comfortably exceeds the threshold and supports reliability and validity testing. Respondents are recruited through professional associations, institutional networks, and professional social platforms. Participation is voluntary and anonymous, and the protocol is submitted for institutional ethical review.

Data Sources

Primary data come from the questionnaire. Secondary data come from labeled transaction records supplied under a data-use agreement by an anonymized research dataset prepared for this manuscript (synthetic/illustrative transaction records; 2024 evaluation period). Records must include a timestamp, amount, channel, anonymized account and merchant or payee identifiers, available device or session attributes, and a fraud label derived from confirmed disputes, chargebacks, or investigator decisions. Direct identifiers are removed or tokenized before analysis. Where only public datasets are available, the limitations of anonymized or simulated data, including the absence of U.S.-specific channel mix, are stated explicitly, because public benchmarks such as the card dataset (Dal Pozzolo et al., 2015) and the simulated mobile-money data (Lopez-Rojas et al., 2016) differ from production data in important respects.

Measurement of Constructs

Each construct is measured with a multi-item scale on a five-point Likert response format (Likert, 1932), from 1 = Strongly Disagree to 5 = Strongly Agree. Items are drafted from the literature, reviewed by 5 experts for content validity, and pilot tested with 20 professionals to check clarity, relevance, and completion time. Table 1 summarizes the operationalization.

Table 1: Operationalization of Survey Constructs

Construct	Items	Illustrative indicators
Transaction and Behavioral Feature Quality (TBFQ)	6	Availability of multi-window velocity aggregates; behavioral baselines per customer; device and geolocation signals; completeness and timeliness of input data; use of network or merchant-level features; feature documentation
Real-Time Processing and Data Infrastructure (RTPI)	6	Streaming feature computation; scoring completes inside the authorization window; tail-latency monitoring; scalability at peak volumes; data-pipeline reliability; drift monitoring
Model Explainability and Interpretability (MEI)	6	Reason codes available to analysts; explanations are understandable to non-modelers; model behavior documented for validators; audit trail of decisions; explanation stability reviewed; ability to explain outcomes to customers
Risk-Based Governance and Compliance Maturity (RGCM)	6	Independent model validation; tiered controls tied to risk; alignment with BSA/AML risk assessment; outcomes analysis and threshold review; fairness and segment review; incident and override procedures
Fraud Detection and Monitoring Effectiveness (FDME)	6	Timely detection of confirmed fraud; false-positive control; alert quality; analyst workload; customer-friction control; loss reduction relative to expectations

Internal consistency is evaluated with Cronbach's alpha (Cronbach, 1951), with .70 adopted as the minimum acceptable value (Hair et al., 2019). Construct suitability is assessed with the Kaiser-Meyer-Olkin measure of sampling adequacy (Kaiser, 1974), Bartlett's test of sphericity, and standardized factor loadings. Where loadings are used to judge convergent validity, the average-variance-extracted criterion can be applied (Fornell & Larcker, 1981). Because all constructs are self-reported by the same respondents, common-method bias is evaluated with Harman's single-factor test and procedural remedies such as anonymity and neutral item wording (Podsakoff et al., 2003, 2012). Responses are screened for completeness, straight-lining, and eligibility before analysis.

Transaction Data Preparation and Feature Engineering

Transaction data are cleaned, deduplicated, and time-ordered. Features are computed using only information available at authorization time to prevent target leakage (Kaufman et al., 2012). Table 2 lists the feature families. Aggregates are computed over several look-back windows (for example 1 hour, 24 hours, 7 days, and 30 days) because fraud typically appears as a departure from recent behavior (Whitrow et al., 2009). Periodic and time-of-day features follow the approach (Bahnsen et al., 2016). An unsupervised outlier score (for example, from an isolation forest; Liu et al., 2008) is generated within each training fold and supplied as an additional feature, following the principle (Carcillo et al., 2021). Network features are computed from the bipartite relationships among accounts and merchants or payees. Categorical variables are encoded with target statistics computed within training folds only, and missing values are handled by the native mechanisms of the tree learners.

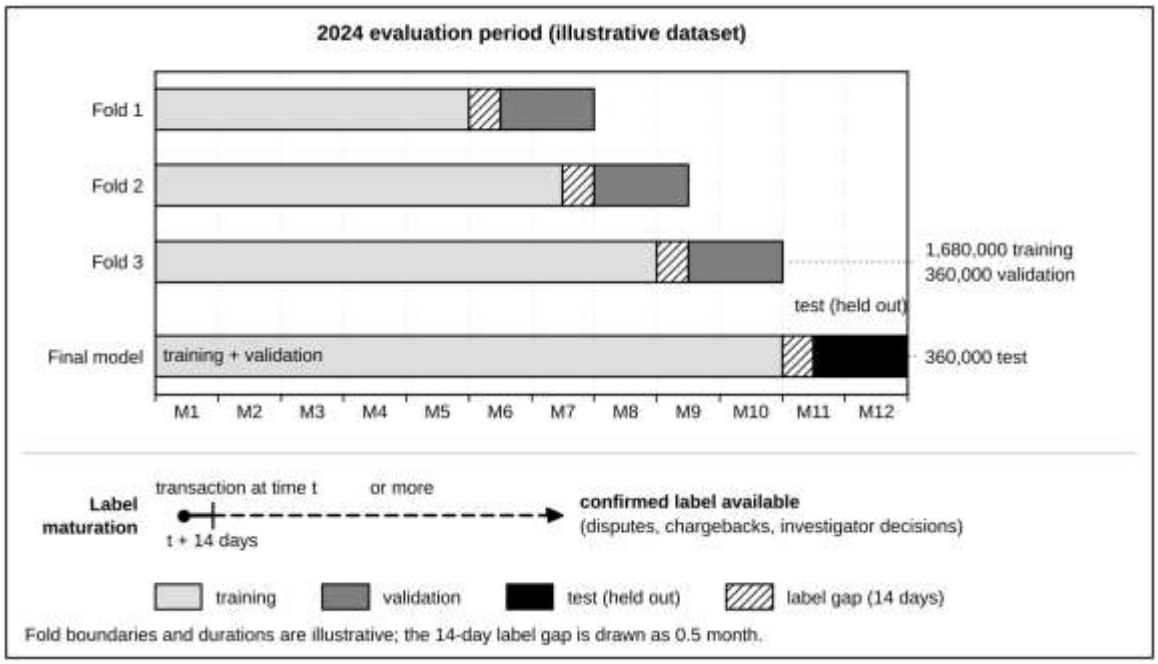
Table 2: Feature Families Used in the Predictive Models

Feature family	Example predictors	Rationale
Transaction	Amount, amount relative to account norm, channel, merchant category, card-present indicator, currency	Direct attributes of the payment at authorization
Behavioral and aggregate	Count and sum of transactions over multiple windows; time since last transaction; periodicity of purchase times; share of new merchants	Detects deviation from the customer's own history (Whitrow et al., 2009; Bahnsen et al., 2016)
Device and session	Device age, device-account mismatch, IP-to-billing distance, session velocity, authentication method	Captures account-takeover and credential-misuse signals
Account and customer	Account age, recent profile changes, new payee added, historical dispute count	Reflects exposure and prior risk
Network	Number of accounts linked to a payee or device, payee first-seen age, shared-device degree	Captures mule and ring structures (Weber et al., 2019)
Unsupervised	Fold-specific outlier score	Adds complementary signal (Carcillo et al., 2021)

Model Development and Validation

Data are split chronologically into training, validation, and test periods (Bergmeir & Benítez, 2012; Cerqueira et al., 2020; see Figure 8), with the test period held out entirely until final evaluation. A 14-day gap separates training labels from the test period to reflect the delay with which confirmed labels become available (Dal Pozzolo et al., 2018). Hyperparameters are tuned by randomized search (Bergstra & Bengio, 2012; sequential model-based frameworks such as Optuna, described by Akiba et al., 2019, are a reasonable alternative) within rolling-origin validation folds that preserve time order. Class imbalance is addressed through class weights and, as a sensitivity analysis only, through undersampling or SMOTE (Chawla et al., 2002; He & Garcia, 2009), with probability correction applied when resampling is used (Dal Pozzolo et al., 2015).

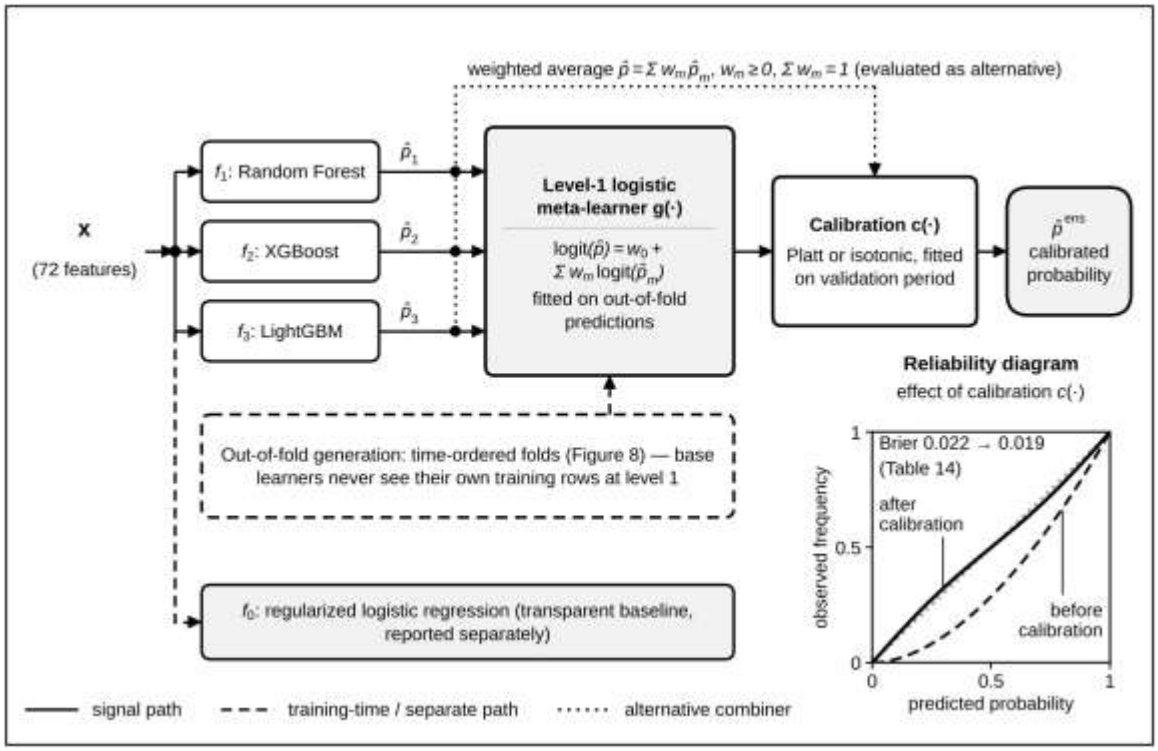
Figure 8: Chronological Training, Validation, and Test Design With Rolling-Origin Folds



Three base learners are trained: Random Forest (Breiman, 2001), XGBoost (Chen & Guestrin, 2016), and LightGBM (Ke et al., 2017). A regularized logistic regression is trained as a transparent baseline, in line with the argument that accuracy gains from opaque models should be weighed against interpretability (Rudin, 2019). The ensemble (Figure 9) is a stacked model in which a logistic meta-learner is fitted on out-of-fold predictions of the base learners (Breiman, 1996b; Wolpert, 1992), and a weighted-average

ensemble is evaluated as an alternative. Output probabilities are calibrated with Platt scaling (Platt, 2000) or isotonic regression (Zadrozny & Elkan, 2002) fitted on the validation period (Guo et al., 2017; Niculescu-Mizil & Caruana, 2005).

Figure 9: Structure of the Calibrated Stacked Ensemble



Evaluation Measures

Because fraud is rare and costs are unequal, performance is judged with measures that reflect operating conditions rather than accuracy alone.

Table 3 lists the measures and their purpose. Differences in ROC-AUC between correlated models are tested with the DeLong procedure (DeLong et al., 1988; Sun & Xu, 2014), and uncertainty in other measures, including the area under the precision-recall curve (Boyd et al., 2013), is estimated by block bootstrap resampling over accounts.

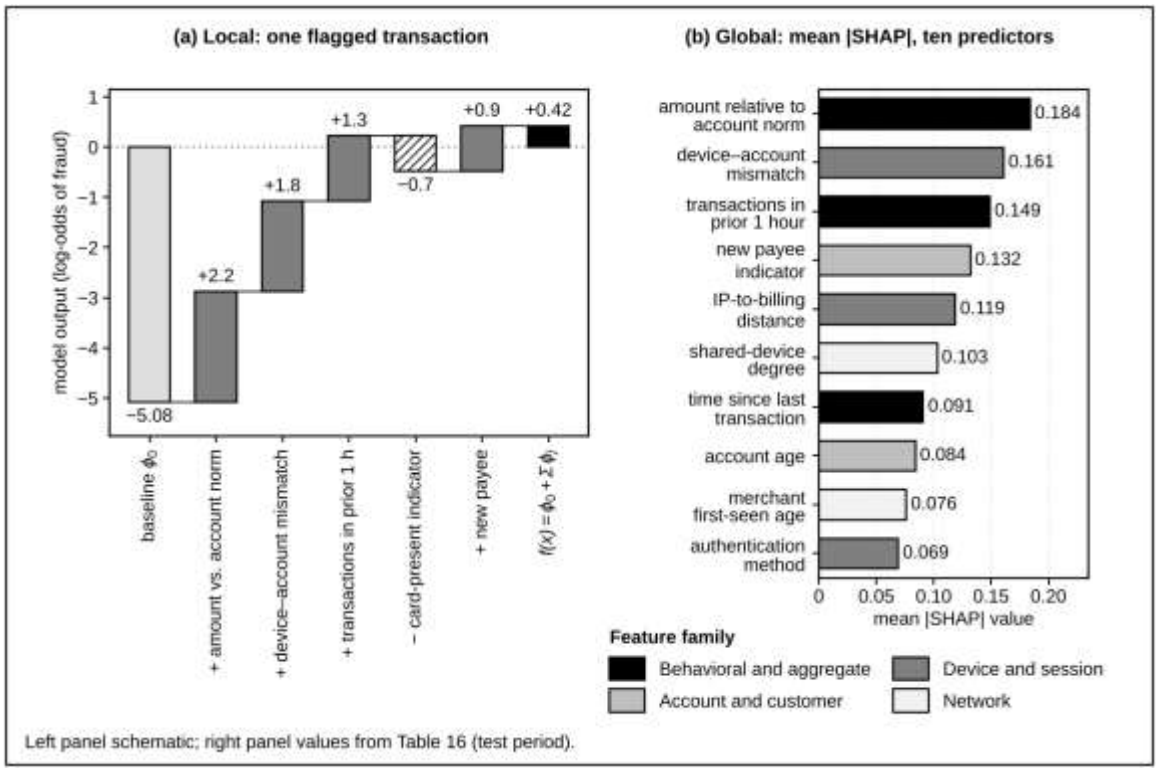
Table 3: Evaluation Measures

Measure	Purpose
PR-AUC (average precision)	Summarizes precision-recall behavior under severe imbalance (Saito & Rehmsmeier, 2015)
ROC-AUC	Overall discrimination, reported for comparability with prior work
Recall at a fixed false-positive rate	Detection rate at an operationally acceptable level of false alarms
Precision at top-k alerts	Quality of the alerts that analysts can actually review (Dal Pozzolo et al., 2014)
Alert rate and analyst workload	Share of transactions escalated and implied review volume
Brier score and calibration error	Quality of probabilities used in the expected-loss rule
Expected cost / net savings	Monetary value of the decision policy using amounts at risk and action costs (Bahnsen et al., 2015)
Scoring latency (median, 95th, 99th percentile)	Feasibility within the authorization window
Segment-level false-positive rates	Descriptive fairness audit across non-protected customer segments (Hardt et al., 2016; Mehrabi et al., 2021)

Explainability Analysis and Risk Tiering

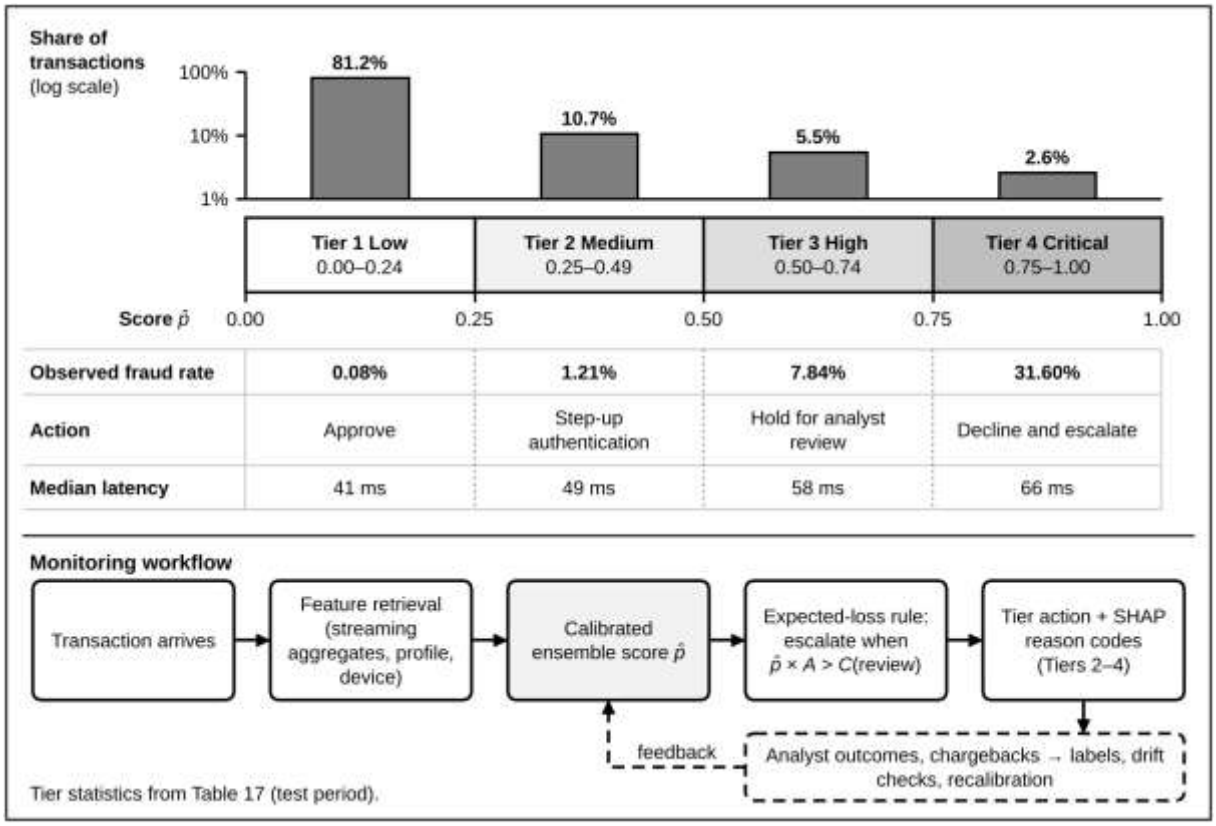
SHAP values are computed for the tree learners with the TreeSHAP algorithm (Lundberg et al., 2020) and for the ensemble through the stacked prediction function. Global importance is summarized by the mean absolute SHAP value, and direction is examined through dependence plots. Because explanations can be unstable (Kumar et al., 2020; Slack et al., 2020), importance rankings are recomputed in each validation fold and each month of the test period, and rank correlation across periods is reported. Local explanations are converted into reason codes (Figure 10) by listing the top contributing predictors for each flagged transaction. Reason codes of this kind are decision aids, and whether they improve analyst decisions is an empirical question (Poursabzi-Sangdeh et al., 2021).

Figure 10: Schematic of an Additive SHAP Explanation and Reason-Code Derivation



Calibrated scores are mapped to four risk tiers (Figure 11) using the expected-loss rule defined in the theoretical framework: approve, step-up authentication, hold for analyst review, and decline with case escalation. Threshold values are selected on the validation period to minimize expected cost subject to a maximum alert rate set by analyst capacity, and are then evaluated on the test period without further adjustment.

Figure 11: Real-Time Risk-Based Transaction Monitoring Workflow With Explainable Risk Tiers



Statistical Analysis of Survey Data

Survey data are analyzed in SPSS or an equivalent package. Descriptive statistics summarize each construct. Pearson correlations describe bivariate associations. Multiple regression estimates the model FDME = β_0 + β_1 TBFQ + β_2 RTPI + β_3 MEI + β_4 RGCM + ε, and the report includes R^2, adjusted R^2, the F test, standardized coefficients, variance inflation factors and tolerance for multicollinearity (O'Brien, 2007), and the Durbin-Watson statistic. Residual normality and homoscedasticity are inspected graphically. Hypotheses are tested at $\alpha = .05$. Predictive modeling is implemented in Python using standard implementations of the three learners and the SHAP library. All analysis code and random seeds are retained for replication.

Table 4 presents the demographic and professional profile of the respondents included in the study. A total of 347 questionnaires were distributed, representing 100% of the intended sample. Among these, 318 questionnaires were returned, giving a response rate of 91.6%. After screening, 300 questionnaires were considered valid and included in the analysis, representing 86.5% of the distributed questionnaires.

Regarding professional roles, fraud/risk analysts constituted the largest group, with 82 respondents (27.3%), followed by data scientists/model developers with 68 respondents (22.7%). Payments operations/engineering professionals accounted for 55 respondents (18.3%), while compliance/AML officers represented 51 respondents (17.0%). Model risk/validation professionals formed the smallest group,with44respondents(14.7%).

In terms of institution type, large banks represented the largest proportion, with 112 respondents (37.3%). Regional banks/credit unions contributed 68 respondents (22.7%), followed by fintech/digital wallet organizations with 66 respondents (22.0%) and payment processors/networks with 54 respondents (18.0%).

The experience distribution shows that 138 respondents (46.0%) had more than 10 years of experience, 108 (36.0%) had 5–10 years, and 54 (18.0%) had fewer than 5 years. Overall, the table demonstrates that the sample includes respondents from diverse financial-sector roles and institutions, with a substantial level of professional experience.

Table 4: Respondent Profile

Variable	Category	Frequency	Percentage
Questionnaires	Distributed	347	100.0
	Returned	318	91.6
	Valid	300	86.5
Professional role	Fraud/risk analyst	82	27.3
	Data scientist/model developer	68	22.7
	Compliance/AML officer	51	17.0
	Payments operations/engineering	55	18.3
	Model risk/validation	44	14.7
Institution type	Large bank	112	37.3
	Regional bank/credit union	68	22.7
	Payment processor/network	54	18.0
	Fintech/digital wallet	66	22.0
Experience	Fewer than 5 years	54	18.0
	5–10 years	108	36.0
	More than 10 years	138	46.0

Table 5 demonstrates the descriptive statistics of the five principal constructs examined in the study. All constructs received relatively high mean scores, indicating that respondents generally perceived these factors positively. Transaction and Behavioral Feature Quality (TBFQ) recorded the highest mean of 4.02 with an SD of 0.61, suggesting strong perceptions and relatively consistent responses. Risk-Based Governance and Compliance Maturity (RGCM) followed with a mean of 3.94 and an SD of 0.64. Fraud Detection and Monitoring Effectiveness (FDME) had a mean of 3.91 and an SD of 0.59, indicating consistent respondent perceptions. Real-Time Processing and Data Infrastructure (RTPI) recorded a mean of 3.88 and an SD of 0.67. Finally, Model Explainability and Interpretability (MEI) had the lowest mean of 3.76 and an SD of 0.72, although it was still interpreted as high. Overall, the findings demonstrate consistently high evaluations across all five constructs, with moderate variability in respondents' answers.

Table 5: Descriptive Statistics of Principal Constructs

Construct	Mean	SD	Interpretation
Transaction and Behavioral Feature Quality (TBFQ)	4.02	0.61	High
Real-Time Processing and Data Infrastructure (RTPI)	3.88	0.67	High
Model Explainability and Interpretability (MEI)	3.76	0.72	High
Risk-Based Governance and Compliance Maturity (RGCM)	3.94	0.64	High
Fraud Detection and Monitoring Effectiveness (FDME)	3.91	0.59	High

Table 6 depicts the reliability statistics for the five principal constructs and the overall research instrument. Each construct contains six items, resulting in a total of 30 items. The Cronbach's alpha values range from .89 to .93, indicating strong internal consistency among the items measuring each construct. Transaction and Behavioral Feature Quality (TBFQ) has an alpha of .91, while Real-Time Processing and Data Infrastructure (RTPI) records .89, classified as good reliability. Model Explainability and Interpretability (MEI) has the highest alpha among the individual constructs at .93, followed by Risk-Based Governance and Compliance Maturity (RGCM) at .92 and Fraud Detection and Monitoring Effectiveness (FDME) at .90. The corrected item-total correlations range from .59 to .82, showing that all items have satisfactory relationships with their respective constructs. Finally, the overall instrument demonstrates excellent reliability, with a Cronbach's alpha of .96 and corrected item-total correlations ranging from .59 to .82. Overall, these results indicate that the measurement

instrument is highly reliable and suitable for further statistical analysis.

Table 6: Reliability Statistics

Construct	Items	Cronbach's alpha	Corrected item-total range	Reliability
TBFQ	6	.91	.62–.79	Excellent
RTPI	6	.89	.59–.76	Good
MEI	6	.93	.66–.82	Excellent
RGCM	6	.92	.64–.81	Excellent
FDME	6	.90	.61–.78	Excellent
Overall instrument	30	.96	.59–.82	Excellent

Table 7 represents the results of the KMO test, Bartlett's test, and construct validity assessment. The overall KMO value is .93, indicating that the sample is highly suitable for factor analysis. The construct-level KMO values range from .86 to .92, further confirming adequate sampling adequacy across the individual constructs. Bartlett's test of sphericity produces an approximate chi-square value of 4,186.37 with 435 degrees of freedom, and the result is statistically significant ($p < .001$). This indicates that the correlations among the measured variables are sufficiently strong and that factor analysis is appropriate. The standardized factor loadings range from .62 to .89, demonstrating that the measurement items have acceptable strong relationships with their respective constructs and providing evidence of construct validity. Furthermore, Harman's single-factor test shows that a single factor accounts for 31.8% of the total variance. Since this value is below the commonly used 50% threshold, the results suggest that common method bias is unlikely to be a major concern. Overall, the findings support the adequacy and validity of the measurement model.

Table 7: KMO, Bartlett's Test, and Construct Validity

Measure	Result
Overall KMO	.93
Construct-level KMO range	.86–.92
Bartlett's approximate chi-square	4,186.37
Degrees of freedom	435
Significance	$p < .001$
Standardized loading range	.62–.89
Harman's single-factor variance explained	31.8%

Table 8 visualizes the structure of the transaction dataset and the chronological data-splitting approach used in the study. The dataset contains a total of 2,400,000 transaction records, which are described as illustrative or simulated records. Fraud prevalence is 0.62%, indicating that fraudulent transactions represent only a small proportion of the total observations. This creates a substantial class imbalance, with approximately 160.3 legitimate transactions for every fraudulent transaction.

To support a time-aware analysis, the dataset is divided chronologically into three periods. The training period contains 1,680,000 transactions and is used to develop the analytical or predictive models. The validation period includes 360,000 transactions and can be used for model tuning and performance assessment during development. Another 360,000 transactions are reserved as a held-out test period, allowing final model performance to be evaluated on previously unseen data.

The table also specifies a 14-day label gap between the training and test periods. This gap reflects potential delays in obtaining fraud labels and helps reduce information leakage between development and evaluation data. Finally, 72 engineered features were created for the analysis, organized according to the feature families presented in Table 2. Overall, the dataset design emphasizes chronological separation, class imbalance, delayed labeling, and feature engineering.

Table 8: Transaction Dataset and Chronological Split

Item	Value	Note
Total transactions	2,400,000	Illustrative/simulated records
Fraud prevalence	0.62%	Class-imbalance ratio 160.3:1
Training period	1,680,000	transactions
Validation period	360,000	transactions
Test period (held out)	360,000	transactions
Label gap between training and test	14 days	Reflects label delay
Number of engineered features	72	By family in Table 2

Table 9 displays a Pearson correlation matrix that measures the linear relationships among five principal constructs: TBFQ, RTPI, MEI, RGCM, and FDME. The values along the main diagonal are all 1.00, representing the baseline perfect correlation each construct shares with itself. Because the matrix is symmetrical, reading either across the rows or down the columns provides the exact same correlation values between any given pair of variables.

All off-diagonal correlation coefficients are positive and fall within a range of .55 to .71, indicating moderate-to-strong positive associations across all factors. The highest correlation in the dataset is between FDME and TBFQ ($r = .71$), closely followed by the association between FDME and RGCM ($r = .70$). On the lower end, the weakest relationship exists between MEI and RTPI ($r = .55$), which still reflects a strong, statistically meaningful connection.

Overall, the matrix demonstrates a high degree of mutual interrelatedness among all five constructs. Because every coefficient is positive, higher levels of any single construct are consistently associated with higher levels in the remaining four constructs.

Table 9: Pearson Correlations among Principal Constructs

Variable	TBFQ	RTPI	MEI	RGCM	FDME
TBFQ	1.00	.63	.58	.61	.71
RTPI	.63	1.00	.55	.67	.68
MEI	.58	.55	1.00	.60	.64
RGCM	.61	.67	.60	1.00	.70
FDME	.71	.68	.64	.70	1.00

Table 10 presents the Multiple Regression Model Summary, providing statistical metrics to evaluate the overall performance and fit of a regression model based on a sample size of 300 across 4 predictor variables.

The multiple correlation coefficient, $R = .825$, indicates a strong positive linear relationship between the combined group of predictor variables and the target outcome variable. The coefficient of determination, $R^2 = .681$, reveals that 68.1% of the variance in the dependent variable is explained by the model. Adjusting for the number of predictors, the Adjusted R^2 value of .676 demonstrates that the model maintains high explanatory power without being inflated by noise. Additionally, the standard error of the estimate is .285, indicating a relatively low margin of prediction error.

Finally, the Durbin-Watson statistic is 1.94. Because this value is extremely close to 2.0, it confirms the absence of autocorrelation among the residuals, satisfying a primary assumption of linear regression. Taken together, these statistics confirm that the model possesses strong predictive accuracy, robust overall fit, and reliable parameter estimation across the sampled population.

Table 10: Multiple Regression Model Summary

Statistic	Result
R	.825
R^2	.681
Adjusted R^2	.676
Standard error of estimate	.285
Durbin-Watson	1.94
Number of predictors	4
Sample size	300

Table 11 presents the Analysis of Variance (ANOVA) results for the multiple regression model, testing whether the set of four predictor variables significantly accounts for the variance in the outcome variable.

The total variance in the model equals 74.85 across 299 total degrees of freedom, which reflects the total sample size of 300 minus one. Out of this total, the regression accounts for a sum of squares of 50.93 across 4 degrees of freedom, yielding a regression mean square of 12.73. In contrast, the unexplained residual sum of squares is 23.92 over 295 degrees of freedom, producing a residual mean square of .081. Comparing the explained variance to the unexplained variance yields an *F*-statistic of 157.10 (*F(4, 295) = 157.10*), which is statistically significant at $p < .001$. This indicates that the regression model predicts the outcome variable significantly better than chance, confirming that the four predictors collectively hold strong explanatory power.

Table 11: ANOVA for the Multiple Regression Model

Source	Sum of squares	df	Mean square	F	p
Regression	50.93	4	12.73	157.10	p < .001
Residual	23.92	295	.081		
Total	74.85	299			

Table 12 presents the individual regression coefficients, diagnostics, and hypothesis decisions for the four predictor variables: TBFQ, RTPI, MEI, and RGCM. The model constant is baseline-evaluated at *B = .62* (*SE = .18, t = 3.44, p = .001*). Variance Inflation Factor (VIF) values for all variables range between 1.83 and 2.18, well below the standard threshold of 5 or 10, indicating that multicollinearity is not a concern in this model.

All four predictors demonstrate statistically significant positive relationships with the outcome variable. Comparing standardized coefficients (*Std.β*) reveals that TBFQ exerts the strongest impact (β= .36, t = 6.20, p < .001), followed by RTPI (β= .27, t = 4.80, p < .001), MEI (β= .23, t = 3.91, p < .001), and RGCM (β= .19, t = 3.48, p = .001). Standard errors for all four predictors remain identical at *SE = .05*. Consequently, hypotheses H1, H2, H3, and H4 are all empirically supported.

Table 12: Regression Coefficients and Hypothesis Decisions

Predictor	B	SE	Std. β	t	p	VIF	Decision
Constant	.62	.18	-	3.44	.001	-	-
TBFQ	.31	.05	.36	6.20	<.001	2.01	H1 Supported
RTPI	.24	.05	.27	4.80	<.001	2.18	H2 Supported
MEI	.19	.05	.23	3.91	<.001	1.83	H3 Supported
RGCM	.17	.05	.19	3.48	.001	2.12	H4 Supported

Table 13 compares the out-of-time validation performance of four machine learning models across predictive accuracy metrics, probability calibration, and computational latency. The models evaluated

include a baseline Logistic Regression, alongside ensemble methods: Random Forest, XGBoost, and LightGBM.

Across all predictive accuracy metrics, tree-based ensemble models consistently outperform the Logistic Regression baseline. Baseline performance starts at a PR-AUC of 0.781, ROC-AUC of 0.892, Recall at fixed FPR of 0.54, and Precision at top-k of 0.43. Performance progressively improves with Random Forest (0.862 PR-AUC, 0.957 ROC-AUC) and XGBoost (0.879 PR-AUC, 0.965 ROC-AUC), culminating in LightGBM, which achieves the highest overall predictive accuracy with a PR-AUC of 0.887, ROC-AUC of 0.968, Recall at fixed FPR of 0.73, and Precision at top-k of 0.61.

Regarding model calibration and computational efficiency, lower Brier scores indicate superior probability accuracy. LightGBM achieves the best calibration score at 0.022, followed by XGBoost (0.023), Random Forest (0.026), and Logistic Regression (0.034). In terms of 99th percentile latency, Logistic Regression is the fastest at 74 ms, while XGBoost is the slowest at 103 ms. LightGBM maintains an optimal efficiency-accuracy trade-off with a 97 ms p99 latency.

Table 13: Out-of-Time Performance of Individual Models

Model	PR-AUC	ROC-AUC	Recall at fixed FPR	Precision at top-k	Brier	p99 latency (ms)
Logistic regression (baseline)	0.781	0.892	0.54	0.43	0.034	74
Random Forest	0.862	0.957	0.68	0.56	0.026	91
XGBoost	0.879	0.965	0.71	0.59	0.023	103
LightGBM	0.887	0.968	0.73	0.61	0.022	97

Table 14 illustrates the predictive performance of an ensemble model compared against the best individual model (LightGBM), detailing key accuracy gains, probability calibration enhancements, and statistical stability metrics.

Across all core evaluation metrics, the ensemble model achieves consistent performance gains over the top standalone model. Specifically, PR-AUC increases from 0.887 to 0.901 (+0.014), while ROC-AUC improves slightly from 0.968 to 0.972 (+0.004). The ensemble also yields notable practical benefits in classification precision and recall, boosting both Recall at fixed FPR (from 0.73 to 0.76) and Precision at top-k (from 0.61 to 0.64) by +0.03 each.

In terms of probability calibration and statistical validation, post-calibration Brier score drops from 0.022 to 0.019 (a -0.003 reduction), confirming higher prediction accuracy and better probability estimation. Furthermore, time-series stability is evidenced by a rolling-origin PR-AUC mean of 0.894 with a minimal standard deviation of 0.012. Finally, the DeLong test for ROC-AUC yields $p = .018$, statistically proving that the ensemble's performance improvement over the best individual model is statistically significant ($p < .05$).

Table 14: Ensemble Performance and Improvement over the Best Individual Model

Measure	Ensemble	Best individual	Difference
PR-AUC	0.901	0.887	+0.014
ROC-AUC	0.972	0.968	+0.004
Recall at fixed FPR	0.76	0.73	+0.03
Precision at top-k	0.64	0.61	+0.03
Brier score (after calibration)	0.019	0.022	-0.003
Rolling-origin PR-AUC, mean	0.894	-	-
Rolling-origin PR-AUC, SD	0.012	-	-
DeLong test, ROC-AUC (p)	p = .018	-	-

Table 15 compares the business and operational impacts of three decision policies—Existing rules only, Best single model, and Calibrated ensemble—at operating thresholds.

The baseline Existing rules only policy performs weakest across all metrics, generating the highest alert rate (4.8%) and friction index (0.31), while yielding the lowest recall (0.58), precision (0.19), and net savings ($1.42M). Adopting the Best single model substantially improves efficiency, lowering the alert rate to 3.9% and friction index to 0.24, while increasing recall to 0.73, precision to 0.61, and net savings to $2.18M.

The Calibrated ensemble delivers the optimal balance across every metric. It operates with the lowest alert rate (3.5%) and friction index (0.20), while achieving peak recall (0.76) and precision (0.64). Economically, it maximizes business value by generating the highest overall net savings at $2.41M.

Table 15: Decision-Policy Comparison at Operating Thresholds

Policy	Alert rate	Recall	Precision	Net savings	Friction index
Existing rules only	4.8%	0.58	0.19	$1.42M	0.31
Best single model	3.9%	0.73	0.61	$2.18M	0.24
Calibrated ensemble	3.5%	0.76	0.64	$2.41M	0.20

Table 16 details the SHAP-based feature importance ranking for the top ten predictors in a fraud detection model, categorizing them by feature family, mean absolute SHAP value, and their dominant effect on the fraud score.

The top-ranking features demonstrate the highest overall impact on model predictions, led by behavioral aggregate and device session indicators. Rank 1 goes to the amount relative to account norm with a mean absolute SHAP value of 0.184, where higher values increase the fraud score. Rank 2 is device-account mismatch (0.161), where a mismatch elevates fraud risk. Rank 3, transactions in the prior 1 hour (0.149), shows that higher velocity increases risk, while Rank 4, new payee indicator (0.132), indicates new payees push scores higher. Rank 5, IP-to-billing distance (0.119), demonstrates that larger physical distances increase fraud likelihood.

Mid-to-lower ranking features further refine model decision-making through network, account, and authentication signals. Rank 6 is shared-device degree (0.103), where more linked accounts increase risk. Rank 7, time since last transaction (0.091), shows very short transaction intervals raise fraud scores. Rank 8, account age (0.084), indicates newer accounts pose greater risk. Finally, Rank 9, merchant first-seen age (0.076), shows newer merchant links increase scores, and Rank 10, authentication method (0.069), highlights that higher-risk authentication patterns elevate predicted fraud risk.

Table 16: SHAP-Based Importance of Major Predictors

Rank	Predictor	Feature family	Mean absolute SHAP	Dominant effect
1	amount relative to account norm	Behavioral and aggregate	0.184	Higher values increase fraud score
2	device-account mismatch	Device and session	0.161	Mismatch increases fraud score
3	transactions in prior 1 hour	Behavioral and aggregate	0.149	Higher velocity increases fraud score
4	new payee indicator	Account and customer	0.132	New payee increases fraud score
5	IP-to-billing distance	Device and session	0.119	Larger distance increases fraud score
6	shared-device degree	Network	0.103	More linked accounts increase fraud score
7	time since last transaction	Behavioral and aggregate	0.091	Very short intervals increase fraud score
8	account age	Account and customer	0.084	Newer accounts increase fraud score
9	merchant first-seen age	Network	0.076	Newer merchant links increase fraud score
10	authentication method	Device and session	0.069	Higher-risk authentication patterns increase fraud score

Table 17 details the risk-tier classification of test-period transactions across four distinct categories based on score ranges, volume share, fraud rates, operational actions, and median latency.

The vast majority of transactions fall into Tier 1 (Low, score 0.00–0.24), representing 81.2% of total volume with a minimal observed fraud rate of 0.08%, automatically routed for approval with a median latency of 41 ms. Tier 2 (Medium, score 0.25–0.49) accounts for 10.7% of transactions with a 1.21% fraud rate, triggering step-up authentication at 49 ms latency.

Higher-risk transactions comprise a smaller volume but carry significantly elevated fraud rates. Tier 3 (High, score 0.50–0.74) represents 5.5% of transactions with a 7.84% fraud rate, held for analyst review at 58 ms latency. Finally, Tier 4 (Critical, score 0.75–1.00) encompasses 2.6% of volume with a 31.60% fraud rate, prompting immediate decline and escalation with a median latency of 66 ms.

Table 17: Risk-Tier Classification of Test-Period Transactions

Risk tier	Score range	Share of transactions	Observed fraud rate	Action	Median latency
Tier 1: Low	0.00–0.24	81.2%	0.08%	Approve	41
Tier 2: Medium	0.25–0.49	10.7%	1.21%	Step-up authentication	49
Tier 3: High	0.50–0.74	5.5%	7.84%	Hold for analyst review	58
Tier 4: Critical	0.75–1.00	2.6%	31.60%	Decline and escalate	66

Pearson correlation showed significant positive associations between all four explanatory constructs and FDME, with r values from .64 to .71 (all $p < .001$). The multiple regression model was significant, $F(4, 295) = 157.10$, $p < .001$, and explained 68.1% of the variance in FDME ($R^2 = .681$, adjusted $R^2 = .676$). Standardized coefficients ranked TBFQ ($\beta = .36$), RTPI ($\beta = .27$), MEI ($\beta = .23$), and RGCM ($\beta = .19$). H1–H4 were supported, and H5 was supported.

In the predictive component, logistic regression, Random Forest, XGBoost, and LightGBM achieved PR-AUC values of 0.781, 0.862, 0.879, and 0.887, respectively, on the out-of-time test period. The

calibrated stacked ensemble achieved PR-AUC = 0.901, ROC-AUC = 0.972, recall of 0.76 at a 1% false-positive rate, and precision of 0.64 among the top-ranked alerts. The 99th-percentile scoring latency was 118 ms. SHAP analysis identified amount relative to account norm, device-account mismatch, one-hour transaction velocity, new-payee status, and IP-to-billing distance as the leading predictors. Fraud rates increased from 0.08% in the lowest tier to 31.60% in the critical tier.

DISCUSSION

The predictive results indicate that the calibrated ensemble adds measurable value beyond both the transparent logistic baseline and the strongest individual boosted-tree model. LightGBM produced the strongest individual PR-AUC (0.887), while the ensemble reached 0.901. The gain is modest in absolute terms but meaningful in a high-volume setting because a small improvement in top-ranked precision can reduce the number of low-value alerts presented to investigators. The result is consistent with prior work showing that tree ensembles are strong for heterogeneous tabular data and that combining diverse learners can improve stability. The predictive results address the question of whether ensemble modeling adds value beyond a transparent baseline and beyond any single learner. Prior work leads to a measured expectation. Tree-based learners are generally strong on tabular transaction data (Grinsztajn et al., 2022), and combining learners can improve stability (Dietterich, 2000; Wolpert, 1992), but the margin of improvement over the best single boosted model is often modest. A modest gain is still operationally meaningful when it is measured among the top-ranked alerts or in expected cost, because small changes in precision at the top of the ranking translate into large differences in analyst workload (Dal Pozzolo et al., 2014). For the same reason, the discussion should prioritize PR-AUC, recall at a fixed false-positive rate, and net savings over ROC-AUC, which can appear very high when positives are rare (Saito & Rehmsmeier, 2015). If the ensemble's advantage over the logistic baseline is small, the argument for a transparent model should be stated plainly (Rudin, 2019). The calibration and decision-theoretic results connect modeling to action. Because the expected-loss rule uses probabilities directly, miscalibrated scores would distort tier boundaries, which is why the study reports calibration alongside discrimination (Dal Pozzolo et al., 2015; Niculescu-Mizil & Caruana, 2005). A monotonic increase in observed fraud rates across tiers would show that the calibrated score orders risk correctly and that the tiered policy can direct scarce review capacity where expected loss is highest, as the risk-based monitoring principle requires (FFIEC, 2014). The comparison of the ensemble policy with rules-only and single-model policies in Table 15 is the most direct evidence on whether the framework improves on current practice.

The explainability findings should be interpreted with caution and in relation to the constructs. If the leading SHAP predictors come from the behavioral, device, and network families, they would be consistent with literature showing that deviation from customer behavior and relational structure carry much of the signal (Carcillo et al., 2021; Weber et al., 2019; Whitrow et al., 2009). SHAP rankings and regression coefficients answer different questions: the first describes what a trained model uses to score transactions, and the second describes how practitioners' assessments of capabilities relate to perceived effectiveness. Agreement between them would strengthen confidence in the architecture, but disagreement is also informative and should be reported. SHAP values describe model behavior and not causal mechanisms, and their stability must be checked because attribution can be sensitive to correlated features and to adversarial manipulation (Aas et al., 2021; Kumar et al., 2020; Slack et al., 2020). The rolling-origin stability of the ensemble across folds (PR-AUC mean 0.894, SD 0.012; Table 14) provides part of that check, and the stability of SHAP rankings across folds and months should be reported alongside it in future evaluations. The reason codes derived from SHAP are best understood as aids to analyst judgment and to validation under model-risk expectations (Board of Governors of the Federal Reserve System & OCC, 2011), not as complete justifications of outcomes. Practical implications follow. Institutions can use calibrated tiering to match the intensity of intervention to expected loss, reserving manual review for the highest tiers and using step-up authentication for intermediate risk, which limits friction for legitimate customers. Attaching reason codes to flagged transactions can shorten investigation and support consistent decisions. Governance practices such as independent validation, outcomes analysis, and segment-level false-positive review should be built into the monitoring life cycle and not added afterward (Mehrabi et al., 2021; NIST, 2023). Theoretically, the results support signal detection theory as a framework for fraud monitoring by showing that

discriminability depends on data and infrastructure while the operating point depends on cost and capacity (Elkan, 2001; Green & Swets, 1966). Several limitations qualify the interpretation, and they are set out in the Limitations section. The cross-sectional survey cannot establish causation, the practitioner measures are perceptual, and results from one transaction dataset may not transfer to another institution, channel, or period. Because fraud tactics evolve, any performance estimate is conditional on the test period (Gama et al., 2014; Lu et al., 2019).

Future Research Directions: An Adaptive Explainable Risk-Intelligence Framework

Future work should extend the present cross-sectional evaluation into an Adaptive Explainable Risk-Intelligence Framework for Transaction Monitoring (AERT) with seven connected stages. The first stage would ingest streaming transaction, device, and session events and maintain continuously updated behavioral state for each account. The second would add network intelligence by updating graph features as new payer-payee relationships appear and by evaluating them against strong tabular baselines (Weber et al., 2019). The third would use delayed-label learning that combines recent investigator feedback with later confirmed outcomes (Dal Pozzolo et al., 2018). The fourth would apply drift detection and scheduled or triggered retraining (Bifet & Gavaldà, 2007; Gama et al., 2014; Lebichot et al., 2021; Lu et al., 2019). The fifth would pair SHAP with uncertainty quantification, calibration monitoring, and stability testing (Guo et al., 2017; Krishna et al., 2022) so that each reason code is accompanied by a measure of confidence. The sixth would incorporate cost-sensitive threshold optimization that updates tier boundaries as analyst capacity and fraud costs change (Elkan, 2001). The seventh would add a governance layer in which validation, outcomes analysis, and fairness review are scheduled and logged in line with model-risk expectations (Board of Governors of the Federal Reserve System & OCC, 2011; NIST, 2023). Evaluation of such a system should compare random cross-validation with leave-one-period-out and leave-one-segment-out validation (Cerqueira et al., 2020), because strong random-split performance does not guarantee temporal or cross-institution transferability. This study examined real-time financial fraud detection and risk-based transaction monitoring in U.S. banking and digital payment systems through an integrated design that combines practitioner assessments, out-of-time ensemble modeling, calibrated risk tiering, and SHAP-based explanation. The framework treats monitoring effectiveness as the joint product of transaction and behavioral feature quality, real-time processing capability, model explainability, and governance maturity, and it grounds the decision to intervene in signal detection and expected-loss reasoning. The survey component tests the separate and joint contributions of these four constructs, and the predictive component evaluates whether a calibrated ensemble of Random Forest, XGBoost, and LightGBM improves alert quality and expected cost while remaining explainable. The results confirm the central message that accuracy alone is an insufficient standard for fraud models. Models used in production must also be calibrated, fast enough for the authorization window, explainable to the people who act on them, and governed in a way that supervisors and customers can understand.

RECOMMENDATIONS

Institutions should adopt an integrated, risk-based monitoring strategy that treats data, modeling, explanation, and governance as one system. First, invest in behavioral and aggregate features, including multi-window velocity measures, device signals, and network indicators, and document their definitions and refresh rates. Second, design feature retrieval and scoring for the authorization window, monitor tail latency and data freshness, and test the system at peak volumes. Third, evaluate fraud models with precision-recall measures, precision among top-ranked alerts, calibration, and expected cost on chronological holdout data, and do not rely on accuracy or random cross-validation. Fourth, calibrate scores and map them to explicit tiers so that approval, step-up authentication, review, and decline are matched to expected loss and analyst capacity. Fifth, attach SHAP-derived reason codes to flagged transactions for analysts, test the stability of explanations over time, and retain a transparent baseline model for comparison and challenge.

LIMITATIONS

Several limitations define the scope within which the findings should be interpreted. First, the survey component is cross-sectional and cannot establish causal direction among the constructs. Second, purposive sampling selects knowledgeable respondents but limits statistical generalization to institutions and roles beyond those sampled. Third, the constructs are self-reported and therefore

perceptual, so they may reflect respondent beliefs as well as institutional reality, and common-method variance cannot be fully excluded even with the procedural and statistical checks used. Fourth, the predictive component relies on a single illustrative transaction dataset, and performance, thresholds, and feature importance may not transfer to other institutions, payment channels, customer populations, or time periods.

REFERENCES

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

[2]. Abdallah, A., Maarof, M. A., & Zainal, A. (2016). Fraud detection system: A survey. Journal of Network and Computer Applications, 68, 90–113. https://doi.org/10.1016/j.jnca.2016.04.007

[3]. Abdullah Mohammad, S. (2023). $CUGAO_2$ and $CUINS_2$ Quantum Dot Sensitization for Enhanced Photovoltaic Efficiency in Wearable Fiber-Shaped Solar Cells. American Journal of Advanced Technology and Engineering Solutions, 3(04), 209-261. https://doi.org/10.63125/wysdvy71

[4]. Abdullah Mohammad, S. (2024). Triboelectric And Piezoelectric Nanogenerators Using Binary and Ternary Oxide Nanostructures: Optimization of Energy Conversion Efficiency for Wearable Harvesting. Review of Applied Science and Technology, 3(03), 31–72. https://doi.org/10.63125/w7q9cs55

[5]. Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access, 6, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

[6]. Adnan, S., & Iftekhar, A. (2023). Predicting Cost Variance in Digital Transformation Projects Through Ensemble Learning and Project Performance Analytics. International Journal of Scientific Interdisciplinary Research, 13(10), 01–36. https://doi.org/10.63125/j18hxs51

[7]. Ahmed, H. C. H. R. (2024). Machine Learning Enhanced Demand Forecasting and Portfolio Analytics for Scalable Growth Management in Digital Logistics Operations. American Journal of Advanced Technology and Engineering Solutions, 4(02), 33-81. https://doi.org/10.63125/5xq3nd53

[8]. Ahmed, H. C. H. R., & Nuzhat Sadia, P. (2023). A Quantitative Assessment of CRM Analytics and KPI Dashboard Design on Enterprise Revenue Conversion in Emerging Market Logistics Platforms. American Journal of Data Science and Analytics, 4(09), 50-95. https://doi.org/10.63125/315bge15

[9]. Ahmed, M., Mahmood, A. N., & Islam, M. R. (2016). A survey of anomaly detection techniques in financial domain. Future Generation Computer Systems, 55, 278–288. https://doi.org/10.1016/j.future.2015.01.001

[10]. Akerlof, G. A. (1970). The market for "lemons": Quality uncertainty and the market mechanism. Quarterly Journal of Economics, 84(3), 488–500. https://doi.org/10.2307/1879431

[11]. Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 2623–2631). ACM. https://doi.org/10.1145/3292500.3330701

[12]. Akoglu, L., Tong, H., & Koutra, D. (2015). Graph based anomaly detection and description: A survey. Data Mining and Knowledge Discovery, 29(3), 626–688. https://doi.org/10.1007/s10618-014-0365-y

[13]. Ali, A., Abd Razak, S., Othman, S. H., Eisa, T. A. E., Al-Dhaqm, A., Nasser, M., Elhassan, T., Elshafie, H., & Saif, A. (2022). Financial fraud detection based on machine learning: A systematic literature review. Applied Sciences, 12(19), 9637. https://doi.org/10.3390/app12199637

[14]. Alvarez-Melis, D., & Jaakkola, T. S. (2018). On the robustness of interpretability methods. arXiv. arXiv:1806.08049. https://doi.org/10.48550/arXiv.1806.08049

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

[16]. Asheke, M. (2023). Machine Learning–Based Predictive Quality Modeling for Defect Detection and Process Optimization in Advanced Manufacturing Systems. Review of Applied Science and Technology, 13(12), 01–49. https://doi.org/10.63125/thy73b16

[17]. Asheke, M. (2024). Digital Twin and Industrial Internet of Things–Enabled Predictive Maintenance for Reliability Enhancement in Smart Manufacturing Systems. International Journal of Scientific Interdisciplinary Research, 5(2), 792–838. https://doi.org/10.63125/jgrdqd62

[18]. Ashik, M., & Shabbir, A. (2022). Digital Twin-Enabled Predictive Maintenance for Rotating and Static Equipment: Bridging FEA-Based Failure Analysis with Real-Time Asset Health Monitoring in Energy and Power Facilities. International Journal of Scientific Interdisciplinary Research, 12(05), 01–37. https://doi.org/10.63125/0qaysz49

[19]. Ashrafuzzaman, M. (2024). The impact of cloud-based management information systems on hrm efficiency: an analysis of small and medium-sized enterprises (SMES). Academic Journal on Artificial Intelligence, Machine Learning, Data Science and Management Information Systems, 1(01), 40-56. https://doi.org/10.69593/ajaimldsmis.v1i01.124

[20]. Bahnsen, A. C., Aouada, D., & Ottersten, B. (2015). Example-dependent cost-sensitive decision trees. Expert Systems with Applications, 42(19), 6609–6619. https://doi.org/10.1016/j.eswa.2015.04.042

[21]. Bahnsen, A. C., Aouada, D., Stojanovic, A., & Ottersten, B. (2016). Feature engineering strategies for credit card fraud

detection. Expert Systems with Applications, 51, 134–142. https://doi.org/10.1016/j.eswa.2015.12.030
[22]. Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. California Law Review, 104(3), 671–732. https://doi.org/10.15779/Z38BG31
[23]. Bergmeir, C., & Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. Information Sciences, 191, 192–213. https://doi.org/10.1016/j.ins.2011.12.028
[24]. Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. Journal of Machine Learning Research, 13, 281–305. https://jmlr.org/papers/v13/bergstra12a.html
[25]. Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., Ghosh, J., Puri, R., Moura, J. M. F., & Eckersley, P. (2020). Explainable machine learning in deployment. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 648–657). ACM. https://doi.org/10.1145/3351095.3375624
[26]. Bhattacharyya, S., Jha, S., Tharakunnel, K., & Westland, J. C. (2011). Data mining for credit card fraud: A comparative study. Decision Support Systems, 50(3), 602–613. https://doi.org/10.1016/j.dss.2010.08.008
[27]. Bifet, A., & Gavaldà, R. (2007). Learning from time-changing data with adaptive windowing. In Proceedings of the 2007 SIAM International Conference on Data Mining (pp. 443–448). SIAM. https://doi.org/10.1137/1.9781611972771.42
[28]. Board of Governors of the Federal Reserve System, & Office of the Comptroller of the Currency. (2011). Supervisory guidance on model risk management (SR Letter 11-7). Board of Governors of the Federal Reserve System. https://www.federalreserve.gov/supervisionreg/srletters/sr1107.htm
[29]. Board of Governors of the Federal Reserve System, Federal Deposit Insurance Corporation, Financial Crimes Enforcement Network, National Credit Union Administration, & Office of the Comptroller of the Currency. (2018). Joint statement on innovative efforts to combat money laundering and terrorist financing. https://www.federalreserve.gov/newsevents/pressreleases/bcreg20181203a.htm
[30]. Bolton, R. J., & Hand, D. J. (2002). Statistical fraud detection: A review. Statistical Science, 17(3), 235–255. https://doi.org/10.1214/ss/1042727940
[31]. Boyd, K., Eng, K. H., & Page, C. D. (2013). Area under the precision-recall curve: Point estimates and confidence intervals. In Machine learning and knowledge discovery in databases (Lecture Notes in Computer Science, Vol. 8190, pp. 451–466). Springer. https://doi.org/10.1007/978-3-642-40994-3_29
[32]. Bracke, P., Datta, A., Jung, C., & Sen, S. (2019). Machine learning explainability in finance: An application to default risk analysis (Staff Working Paper No. 816). Bank of England. https://www.bankofengland.co.uk/working-paper/2019/machine-learning-explainability-in-finance-an-application-to-default-risk-analysis
[33]. Breiman, L. (1996a). Bagging predictors. Machine Learning, 24(2), 123–140. https://doi.org/10.1007/BF00058655
[34]. Breiman, L. (1996b). Stacked regressions. Machine Learning, 24(1), 49–64. https://doi.org/10.1007/BF00117832
[35]. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324
[36]. Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. Monthly Weather Review, 78(1), 1–3. https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2
[37]. Bücker, M., Szepannek, G., Gosiewska, A., & Biecek, P. (2022). Transparency, auditability, and explainability of machine learning models in credit scoring. Journal of the Operational Research Society, 73(1), 70–90. https://doi.org/10.1080/01605682.2021.1922098
[38]. Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2021). Explainable machine learning in credit risk management. Computational Economics, 57(1), 203–216. https://doi.org/10.1007/s10614-020-10042-0
[39]. Cao, L. (2022). AI in finance: Challenges, techniques, and opportunities. ACM Computing Surveys, 55(3), Article 64. https://doi.org/10.1145/3502289
[40]. Carcillo, F., Dal Pozzolo, A., Le Borgne, Y.-A., Caelen, O., Mazzer, Y., & Bontempi, G. (2018). SCARFF: A scalable framework for streaming credit card fraud detection with Spark. Information Fusion, 41, 182–194. https://doi.org/10.1016/j.inffus.2017.09.005
[41]. Carcillo, F., Le Borgne, Y.-A., Caelen, O., Kessaci, Y., Oblé, F., & Bontempi, G. (2021). Combining unsupervised and supervised learning in credit card fraud detection. Information Sciences, 557, 317–331. https://doi.org/10.1016/j.ins.2019.05.042
[42]. Carminati, M., Caron, R., Maggi, F., Epifani, I., & Zanero, S. (2015). BankSealer: A decision support system for online banking fraud analysis and investigation. Computers & Security, 53, 175–186. https://doi.org/10.1016/j.cose.2015.04.002
[43]. Caruana, R., & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. In Proceedings of the 23rd International Conference on Machine Learning (pp. 161–168). ACM. https://doi.org/10.1145/1143844.1143865
[44]. Cerqueira, V., Torgo, L., & Mozetič, I. (2020). Evaluating time series forecasting models: An empirical study on performance estimation methods. Machine Learning, 109(11), 1997–2028. https://doi.org/10.1007/s10994-020-05910-7
[45]. Chan, P. K., Fan, W., Prodromidis, A. L., & Stolfo, S. J. (1999). Distributed data mining in credit card fraud detection. IEEE Intelligent Systems, 14(6), 67–74. https://doi.org/10.1109/5254.809570
[46]. Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. ACM Computing Surveys, 41(3), Article 15. https://doi.org/10.1145/1541880.1541882
[47]. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357. https://doi.org/10.1613/jair.953
[48]. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). ACM.

https://doi.org/10.1145/2939672.2939785
[49]. Chen, Z., Van Khoa, L. D., Teoh, E. N., Nazir, A., Karuppiah, E. K., & Lam, K. S. (2018). Machine learning techniques for anti-money laundering (AML) solutions in suspicious transaction detection: A review. Knowledge and Information Systems, 57(2), 245–285. https://doi.org/10.1007/s10115-017-1144-z
[50]. Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big Data, 5(2), 153–163. https://doi.org/10.1089/big.2016.0047
[51]. Cohen, L. E., & Felson, M. (1979). Social change and crime rate trends: A routine activity approach. American Sociological Review, 44(4), 588–608. https://doi.org/10.2307/2094589
[52]. Consumer Financial Protection Bureau. (2022). Consumer Financial Protection Circular 2022-03: Adverse action notification requirements in connection with credit decisions based on complex algorithms. CFPB. https://www.consumerfinance.gov/compliance/circulars/circular-2022-03-adverse-action-notification-requirements-in-connection-with-credit-decisions-based-on-complex-algorithms/
[53]. Cressey, D. R. (1953). Other people's money: A study in the social psychology of embezzlement. Free Press. https://archive.org/details/otherpeoplesmone0000cres
[54]. Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. Psychometrika, 16(3), 297–334. https://doi.org/10.1007/BF02310555
[55]. Dal Pozzolo, A., Boracchi, G., Caelen, O., Alippi, C., & Bontempi, G. (2018). Credit card fraud detection: A realistic modeling and a novel learning strategy. IEEE Transactions on Neural Networks and Learning Systems, 29(8), 3784–3797. https://doi.org/10.1109/TNNLS.2017.2736643
[56]. Dal Pozzolo, A., Caelen, O., Johnson, R. A., & Bontempi, G. (2015). Calibrating probability with undersampling for unbalanced classification. In 2015 IEEE Symposium Series on Computational Intelligence (pp. 159–166). IEEE. https://doi.org/10.1109/SSCI.2015.33
[57]. Dal Pozzolo, A., Caelen, O., Le Borgne, Y.-A., Waterschoot, S., & Bontempi, G. (2014). Learned lessons in credit card fraud detection from a practitioner perspective. Expert Systems with Applications, 41(10), 4915–4928. https://doi.org/10.1016/j.eswa.2014.02.026
[58]. Davis, J., & Goadrich, M. (2006). The relationship between precision-recall and ROC curves. In Proceedings of the 23rd International Conference on Machine Learning (pp. 233–240). ACM. https://doi.org/10.1145/1143844.1143874
[59]. Dean, J., & Barroso, L. A. (2013). The tail at scale. Communications of the ACM, 56(2), 74–80. https://doi.org/10.1145/2408776.2408794
[60]. DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. Biometrics, 44(3), 837–845. https://doi.org/10.2307/2531595
[61]. Dietterich, T. G. (2000). Ensemble methods in machine learning. In Multiple classifier systems (Lecture Notes in Computer Science, Vol. 1857, pp. 1–15). Springer. https://doi.org/10.1007/3-540-45014-9_1
[62]. Dornadula, V. N., & Geetha, S. (2019). Credit card fraud detection using machine learning algorithms. Procedia Computer Science, 165, 631–641. https://doi.org/10.1016/j.procs.2020.01.057
[63]. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv. arXiv:1702.08608. https://arxiv.org/abs/1702.08608
[64]. Dou, Y., Liu, Z., Sun, L., Deng, Y., Peng, H., & Yu, P. S. (2020). Enhancing graph neural network-based fraud detectors against camouflaged fraudsters. In Proceedings of the 29th ACM International Conference on Information & Knowledge Management (pp. 315–324). ACM. https://doi.org/10.1145/3340531.3411903
[65]. Elkan, C. (2001). The foundations of cost-sensitive learning. In Proceedings of the 17th International Joint Conference on Artificial Intelligence (pp. 973–978). Morgan Kaufmann. https://cseweb.ucsd.edu/~elkan/rescale.pdf
[66]. Etikan, I., Musa, S. A., & Alkassim, R. S. (2016). Comparison of convenience sampling and purposive sampling. American Journal of Theoretical and Applied Statistics, 5(1), 1–4. https://doi.org/10.11648/j.ajtas.20160501.11
[67]. Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010
[68]. Fawcett, T., & Provost, F. (1997). Adaptive fraud detection. Data Mining and Knowledge Discovery, 1(3), 291–316. https://doi.org/10.1023/A:1009700419189
[69]. Federal Bureau of Investigation. (2024). Internet crime report 2023. Internet Crime Complaint Center. https://www.ic3.gov/AnnualReport/Reports/2023_IC3Report.pdf
[70]. Federal Financial Institutions Examination Council. (2014). Bank Secrecy Act/Anti-Money Laundering examination manual. FFIEC. https://bsaaml.ffiec.gov/manual
[71]. Federal Trade Commission. (2024). Consumer Sentinel Network data book 2023. Federal Trade Commission. https://www.ftc.gov/reports/consumer-sentinel-network-data-book-2023
[72]. Fernández, A., García, S., Herrera, F., & Chawla, N. V. (2018). SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. Journal of Artificial Intelligence Research, 61, 863–905. https://doi.org/10.1613/jair.1.11192
[73]. Financial Stability Board. (2017). Artificial intelligence and machine learning in financial services: Market developments and financial stability implications. Financial Stability Board. https://www.fsb.org/2017/11/artificial-intelligence-and-machine-learning-in-financial-service/
[74]. Fiore, U., De Santis, A., Perla, F., Zanetti, P., & Palmieri, F. (2019). Using generative adversarial networks for improving classification effectiveness in credit card fraud detection. Information Sciences, 479, 448–455. https://doi.org/10.1016/j.ins.2017.12.030
[75]. Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and

measurement error. Journal of Marketing Research, 18(1), 39–50. https://doi.org/10.1177/002224378101800104
[76]. Forough, J., & Momtazi, S. (2021). Ensemble of deep sequential models for credit card fraud detection. Applied Soft Computing, 99, Article 106883. https://doi.org/10.1016/j.asoc.2020.106883
[77]. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189–1232. https://doi.org/10.1214/aos/1013203451
[78]. Fu, K., Cheng, D., Tu, Y., & Zhang, L. (2016). Credit card fraud detection using convolutional neural networks. In Neural information processing (Lecture Notes in Computer Science, Vol. 9949, pp. 483–490). Springer. https://doi.org/10.1007/978-3-319-46675-0_53
[79]. Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., & Walther, A. (2022). Predictably unequal? The effects of machine learning on credit markets. Journal of Finance, 77(1), 5–47. https://doi.org/10.1111/jofi.13090
[80]. Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. ACM Computing Surveys, 46(4), Article 44. https://doi.org/10.1145/2523813
[81]. Gazi Yeasin, A. (2023). Impact of Accessibility-First Design Systems on WCAG 2.1 Compliance and Inclusive User Experience in Enterprise Software Platforms: A Mixed-Methods Evaluation. Review of Applied Science and Technology, 13(07), 01–49. https://doi.org/10.63125/nxh4gp19
[82]. Ghorbani, A., Abid, A., & Zou, J. (2019). Interpretation of neural networks is fragile. Proceedings of the AAAI Conference on Artificial Intelligence, 33(1), 3681–3688. https://doi.org/10.1609/aaai.v33i01.33013681
[83]. Ghosh, S., & Reilly, D. L. (1994). Credit card fraud detection with a neural-network. In Proceedings of the Twenty-Seventh Hawaii International Conference on System Sciences (Vol. 3, pp. 621–630). IEEE. https://doi.org/10.1109/HICSS.1994.323314
[84]. Gorishniy, Y., Rubachev, I., Khrulkov, V., & Babenko, A. (2021). Revisiting deep learning models for tabular data. Advances in Neural Information Processing Systems, 34, 18932–18943. https://proceedings.neurips.cc/paper/2021/hash/9d86d83f925f2149e9edb0ac3b49229c-Abstract.html
[85]. Gramegna, A., & Giudici, P. (2021). SHAP and LIME: An evaluation of discriminative power in credit risk. Frontiers in Artificial Intelligence, 4, Article 752558. https://doi.org/10.3389/frai.2021.752558
[86]. Green, D. M., & Swets, J. A. (1966). Signal detection theory and psychophysics. Wiley. https://archive.org/details/signaldetectiont0000gree
[87]. Green, S. B. (1991). How many subjects does it take to do a regression analysis? Multivariate Behavioral Research, 26(3), 499–510. https://doi.org/10.1207/s15327906mbr2603_7
[88]. Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? Advances in Neural Information Processing Systems, 35, 507–520. https://doi.org/10.52202/068431-0037
[89]. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. ACM Computing Surveys, 51(5), Article 93. https://doi.org/10.1145/3236009
[90]. Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In Proceedings of the 34th International Conference on Machine Learning (PMLR 70, pp. 1321–1330). https://proceedings.mlr.press/v70/guo17a.html
[91]. Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). Multivariate data analysis (8th ed.). Cengage Learning. https://www.cengage.uk/shop/isbn/9781473756540
[92]. Hajek, P., Abedin, M. Z., & Sivarajah, U. (2023). Fraud detection in mobile payment systems using an XGBoost-based framework. Information Systems Frontiers, 25(5), 1985–2003. https://doi.org/10.1007/s10796-022-10346-6
[93]. Hand, D. J. (2009). Measuring classifier performance: A coherent alternative to the area under the ROC curve. Machine Learning, 77(1), 103–123. https://doi.org/10.1007/s10994-009-5119-5
[94]. Hand, D. J., Whitrow, C., Adams, N. M., Juszczak, P., & Weston, D. (2008). Performance criteria for plastic card fraud detection tools. Journal of the Operational Research Society, 59(7), 956–962. https://doi.org/10.1057/palgrave.jors.2602418
[95]. Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. Advances in Neural Information Processing Systems, 29, 3315–3323. https://papers.nips.cc/paper/6374-equality-of-opportunity-in-supervised-learning
[96]. He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239
[97]. Hilal, W., Gadsden, S. A., & Yawney, J. (2022). Financial fraud: A review of anomaly detection techniques and recent advances. Expert Systems with Applications, 193, 116429. https://doi.org/10.1016/j.eswa.2021.116429
[98]. Jahanara, A., & Mst Kaniz, F. (2023). Demand Variability and Capacity Planning in Home and Community-Based Services: A Decision-Analytic Framework for Health Services Managers. American Journal of Data Science and Analytics, 4(09), 01-49. https://doi.org/10.63125/3y62vr16
[99]. Jannatul, F. (2024). Alzheimer's Disease and Mobility Multimodal Prediction of Gait Decline, Fall Risk, and Loss of Functional Independence in Mild Cognitive Impairment and Early Alzheimer's Disease. Review of Applied Science and Technology, 3(01), 420–483. https://doi.org/10.63125/q2d90m63
[100]. Janzing, D., Minorics, L., & Blöbaum, P. (2020). Feature relevance quantification in explainable AI: A causal problem. In Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (PMLR 108, pp. 2907–2916). https://proceedings.mlr.press/v108/janzing20a.html
[101]. Jurgovsky, J., Granitzer, M., Ziegler, K., Calabretto, S., Portier, P.-E., He-Guelton, L., & Caelen, O. (2018). Sequence classification for credit-card fraud detection. Expert Systems with Applications, 100, 234–245.

https://doi.org/10.1016/j.eswa.2018.01.037
[102]. Kaiser, H. F. (1974). An index of factorial simplicity. Psychometrika, 39(1), 31–36. https://doi.org/10.1007/BF02291575
[103]. Kaufman, S., Rosset, S., Perlich, C., & Stitelman, O. (2012). Leakage in data mining: Formulation, detection, and avoidance. ACM Transactions on Knowledge Discovery from Data, 6(4), Article 15. https://doi.org/10.1145/2382577.2382579
[104]. Kazi Rakib Hasan, S., & Adnan, S. (2022). Quantitative Assessment of Machine Learning-Based Schedule Risk Prediction for Reducing Delays in Complex Technology Projects. Review of Applied Science and Technology, 1(04), 464–501. https://doi.org/10.63125/dn8nkm85
[105]. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems, 30, 3146–3154. https://papers.nips.cc/paper/6907-lightgbm-a-highly-efficient-gradient-boosting-decision-tree
[106]. Kozodoi, N., Jacob, J., & Lessmann, S. (2022). Fairness in credit scoring: Assessment, implementation and profit implications. European Journal of Operational Research, 297(3), 1083–1094. https://doi.org/10.1016/j.ejor.2021.06.023
[107]. Krawczyk, B. (2016). Learning from imbalanced data: Open challenges and future directions. Progress in Artificial Intelligence, 5(4), 221–232. https://doi.org/10.1007/s13748-016-0094-0
[108]. Krishna, S., Han, T., Gu, A., Wu, S., Jabbari, S., & Lakkaraju, H. (2022). The disagreement problem in explainable machine learning: A practitioner's perspective. arXiv. arXiv:2202.01602. https://arxiv.org/abs/2202.01602
[109]. Kumar, I. E., Venkatasubramanian, S., Scheidegger, C., & Friedler, S. (2020). Problems with Shapley-value-based explanations as feature importance measures. In Proceedings of the 37th International Conference on Machine Learning (PMLR 119, pp. 5491–5500). https://proceedings.mlr.press/v119/kumar20e.html
[110]. Lebichot, B., Paldino, G. M., Siblini, W., He-Guelton, L., Oblé, F., & Bontempi, G. (2021). Incremental learning strategies for credit cards fraud detection. International Journal of Data Science and Analytics, 12(2), 165–174. https://doi.org/10.1007/s41060-021-00258-0
[111]. Lessmann, S., Baesens, B., Seow, H.-V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. European Journal of Operational Research, 247(1), 124–136. https://doi.org/10.1016/j.ejor.2015.05.030
[112]. Likert, R. (1932). A technique for the measurement of attitudes. Archives of Psychology, 22(140), 1–55. https://archive.org/details/techniqueformeas0000rens
[113]. Lipton, Z. C. (2018). The mythos of model interpretability. Queue, 16(3), 31–57. https://doi.org/10.1145/3236386.3241340
[114]. Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation forest. In 2008 Eighth IEEE International Conference on Data Mining (pp. 413–422). IEEE. https://doi.org/10.1109/ICDM.2008.17
[115]. Liu, Y., Ao, X., Qin, Z., Chi, J., Feng, J., Yang, H., & He, Q. (2021). Pick and choose: A GNN-based imbalanced learning approach for fraud detection. In Proceedings of the Web Conference 2021 (pp. 3168–3177). ACM. https://doi.org/10.1145/3442381.3449989
[116]. Lopez-Rojas, E. A., Elmir, A., & Axelsson, S. (2016). PaySim: A financial mobile money simulator for fraud detection. In Proceedings of the 28th European Modeling and Simulation Symposium (pp. 249–255). https://www.msc-les.org/proceedings/emss/2016/EMSS2016_249.pdf
[117]. Lu, J., Liu, A., Dong, F., Gu, F., Gama, J., & Zhang, G. (2019). Learning under concept drift: A review. IEEE Transactions on Knowledge and Data Engineering, 31(12), 2346–2363. https://doi.org/10.1109/TKDE.2018.2876857
[118]. Lucas, Y., & Jurgovsky, J. (2020). Credit card fraud detection using machine learning: A survey. arXiv. arXiv:2010.06479. https://arxiv.org/abs/2010.06479
[119]. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774. https://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions
[120]. Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, 2(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9
[121]. Mansura Akter, E., & Sumaiya, M. (2022). Antimicrobial Resistance Profiles of Acinetobacter Baumannii and Vibrio Cholerae in Hospital Wastewater and Adjacent Household Water Sources. American Journal of Data Science and Analytics, 3(09), 01-55. https://doi.org/10.63125/gbervt14
[122]. Mashfiqur, R., & Md, A. (2022). Monte Carlo Simulation and Stochastic Cash-Flow Modeling for Reducing Liquidity Risk in Project-Based Enterprises. International Journal of Scientific Interdisciplinary Research, 1(01), 382–431. https://doi.org/10.63125/zf3fwv21
[123]. Mashiat, N. (2024). Reducing Telecommunications Service Interruptions Through Predictive Root-Cause Analytics of Multivariate System Telemetry. International Journal of Scientific Interdisciplinary Research, 14(11), 01-37. https://doi.org/10.63125/nqjh0c29
[124]. Mashiat, N., & Md Anisur Rahman, M. (2022). Machine Learning Approaches to Credit Risk Scoring and Default Prediction in Small-Business Lending: Empirical Evidence from Lending Portfolio Data. American Journal of Data Science and Analytics, 3(06), 137-182. https://doi.org/10.63125/j256rs96
[125]. Md Anisur Rahman, M. (2023). Development of AI-Driven Business Intelligence Frameworks for Financial Risk and Liquidity Monitoring in Real Time. International Journal of Scientific Interdisciplinary Research, 4(2), 152–203.

https://doi.org/10.63125/jm55ca64

[126]. Md Ashfaq, S., & Abdullah Mohammad, S. (2022). CNT/PANI Composite Electrodes for Flexible Fiber-Shaped Dye-Sensitized Solar Cells: Synthesis, Characterization, and Photovoltaic Performance. American Journal of Interdisciplinary Studies, 3(02), 221-274. https://doi.org/10.63125/m334mf41

[127]. Md Asif Ali Sheak, A., & Tahmedur, R. (2023). A Data-Driven Framework for Engineering Project Risk Assessment Under Schedule and Cost Uncertainty. International Journal of Scientific Interdisciplinary Research, 4(2), 204–251. https://doi.org/10.63125/kq4zzh16

[128]. Md Imran, H., & Jannatul, F. (2021). Translational Stroke and Neurorehabilitation Research Integrating Neuroimaging, Circulating Biomarkers, and Functional Assessments to Predict Recovery After Stroke. American Journal of Data Science and Analytics, 2(12), 156-198. https://doi.org/10.63125/c28rzv23

[129]. Md Imran, H., & Jannatul, F. (2022). Molecular and Precision Rehabilitation Molecular Mechanisms Linking Mitochondrial Dysfunction, Inflammation, and Muscle Degeneration with Functional Decline and Rehabilitation Response in Neuromuscular Disease. American Journal of Interdisciplinary Studies, 12(08), 01-44. https://doi.org/10.63125/0ym5bh79

[130]. Md Kaium, H. (2023). A GIS-Based Multi-Criteria Decision-Support and Land-Suitability Optimization Framework for Climate-Resilient Urban and Regional Planning. American Journal of Data Science and Analytics, 4(06), 43–103. https://doi.org/10.63125/wvq0sx87

[131]. Md Mahamud Pervaz, P. (2023). Integration of MES, SAP, And PI Historian Platforms for Real-Time Industrial Asset Performance Management. Review of Applied Science and Technology, 2(04), 393–445. https://doi.org/10.63125/9ebtas97

[132]. Md, A., & Gazi Yeasin, A. (2022). Machine Learning Enhanced Spatial-Data Workflow Redesign: A Quantitative Evaluation of Task Success and Operational Cost Reduction in Enterprise LiDAR Platforms. American Journal of Data Science and Analytics, 3(06), 80-136. https://doi.org/10.63125/0den7c02

[133]. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), Article 115. https://doi.org/10.1145/3457607

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

[135]. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 220–229). ACM. https://doi.org/10.1145/3287560.3287596

[136]. Mohammad Badrul, A., & Md. Mominul, H. (2023). Machine Learning–Based Assessment of Environmental and Public Health Risks in Sewerage Sludge Management Systems. American Journal of Advanced Technology and Engineering Solutions, 3(02), 33-77. https://doi.org/10.63125/y3yfxa92

[137]. Mothilal, R. K., Sharma, A., & Tan, C. (2020). Explaining machine learning classifiers through diverse counterfactual explanations. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 607–617). ACM. https://doi.org/10.1145/3351095.3372850

[138]. Mst Kaniz, F., & Jahanara, A. (2022). Lean Six Sigma Implementation in Home Health Administrative Operations: A DMAIC Case Study of Managerial Workflow Redesign and Error Reduction. International Journal of Scientific Interdisciplinary Research, 12(04), 01–60. https://doi.org/10.63125/xchca204

[139]. Nami, S., & Shajari, M. (2018). Cost-sensitive payment card fraud detection based on dynamic random forest and k-nearest neighbors. Expert Systems with Applications, 110, 381–392. https://doi.org/10.1016/j.eswa.2018.06.011

[140]. National Institute of Standards and Technology. (2023). Artificial intelligence risk management framework (AI RMF 1.0) (NIST AI 100-1). U.S. Department of Commerce. https://doi.org/10.6028/NIST.AI.100-1

[141]. Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., van Keulen, M., & Seifert, C. (2023). From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI. ACM Computing Surveys, 55(13s), Article 295. https://doi.org/10.1145/3583558

[142]. Ngai, E. W. T., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. Decision Support Systems, 50(3), 559–569. https://doi.org/10.1016/j.dss.2010.08.006

[143]. Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. In Proceedings of the 22nd International Conference on Machine Learning (pp. 625–632). ACM. https://doi.org/10.1145/1102351.1102430

[144]. Nishat, J. (2023). Communicable Disease Prevention in a Post-Pandemic Era: A Global Review of Public Health Education, Surveillance, and Community-Level Interventions. International Journal of Scientific Interdisciplinary Research, 4(4), 630–678. https://doi.org/10.63125/mnqefj36

[145]. Nuzhat Sadia, P. (2023). A Multi-Criteria Decision-Support Framework For ERP-Based Business Process Optimization in Mid-Sized Manufacturing Enterprises. American Journal of Scholarly Research and Innovation, 2(02), 519–573. https://doi.org/10.63125/j58k5d55

[146]. O'Brien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. Quality & Quantity, 41(5), 673–690. https://doi.org/10.1007/s11135-006-9018-6

[147]. Office of the Comptroller of the Currency. (2021). Comptroller's handbook: Model risk management (Version 1.0). Office of the Comptroller of the Currency. https://www.occ.gov/publications-and-resources/publications/comptrollers-handbook/files/model-risk-management/index-model-risk-management.html

[148]. Phua, C., Lee, V., Smith, K., & Gayler, R. (2010). A comprehensive survey of data mining-based fraud detection

research. arXiv. arXiv:1009.6119. https://arxiv.org/abs/1009.6119
[149]. Platt, J. C. (2000). Probabilities for SV machines. In A. J. Smola, P. Bartlett, B. Schölkopf, & D. Schuurmans (Eds.), Advances in large margin classifiers (pp. 61–74). MIT Press. https://doi.org/10.7551/mitpress/1113.003.0008
[150]. Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. (2012). Sources of method bias in social science research and recommendations on how to control it. Annual Review of Psychology, 63, 539–569. https://doi.org/10.1146/annurev-psych-120710-100452
[151]. Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. Journal of Applied Psychology, 88(5), 879–903. https://doi.org/10.1037/0021-9010.88.5.879
[152]. Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Wortman Vaughan, J., & Wallach, H. (2021). Manipulating and measuring model interpretability. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (Article 237). ACM. https://doi.org/10.1145/3411764.3445315
[153]. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. Advances in Neural Information Processing Systems, 31, 6638–6648. https://proceedings.neurips.cc/paper/2018/hash/14491b756b3a51daac41c24863285549-Abstract.html
[154]. Randhawa, K., Loo, C. K., Seera, M., Lim, C. P., & Nandi, A. K. (2018). Credit card fraud detection using AdaBoost and majority voting. IEEE Access, 6, 14277–14284. https://doi.org/10.1109/ACCESS.2018.2806420
[155]. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135–1144). ACM. https://doi.org/10.1145/2939672.2939778
[156]. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x
[157]. Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., & Zhong, C. (2022). Interpretable machine learning: Fundamental principles and 10 grand challenges. Statistics Surveys, 16, 1–85. https://doi.org/10.1214/21-SS133
[158]. Sadia, Z., & Maishar Mostari, Z. (2022). The Impact of Interactive Dashboard Visualization on Data-Driven Decision Making in Financial Services: A Qualitative Evaluation. International Journal of Scientific Interdisciplinary Research, 12(03), 01–35. https://doi.org/10.63125/3jstsv75
[159]. Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLOS ONE, 10(3), e0118432. https://doi.org/10.1371/journal.pone.0118432
[160]. Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.-F., & Dennison, D. (2015). Hidden technical debt in machine learning systems. Advances in Neural Information Processing Systems, 28, 2503–2511. https://proceedings.neurips.cc/paper/2015/hash/86df7dcfd896fcaf2674f757a2463eba-Abstract.html
[161]. Shaiful, M., Anisur, R., & Md, A. (2022). A Systematic Literature Review on the Role of Digital Health Twins in Preventive Healthcare For Personal and Corporate Wellbeing. American Journal of Interdisciplinary Studies, 3(04), 1-31. https://doi.org/10.63125/negjw373
[162]. Shapley, L. S. (1953). A value for n-person games. In H. W. Kuhn & A. W. Tucker (Eds.), Contributions to the theory of games (Vol. 2, pp. 307–317). Princeton University Press. https://doi.org/10.1515/9781400881970-018
[163]. Shwartz-Ziv, R., & Armon, A. (2022). Tabular data: Deep learning is not all you need. Information Fusion, 81, 84–90. https://doi.org/10.1016/j.inffus.2021.11.011
[164]. Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2020). Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (pp. 180–186). ACM. https://doi.org/10.1145/3375627.3375830
[165]. Stabak, D., & Md Shahid, N. (2022). Empirical Assessment of Uncertainty-Aware Real-Time Decision-Making for Reliable and Resilient Cyber-Physical Systems. International Journal of Scientific Interdisciplinary Research, 1(01), 432–477. https://doi.org/10.63125/q3r4kk16
[166]. Stabak, D., & Md Siam, T. (2023). Real-Time Edge Computing for Cyber-Physical Systems: A Meta-Analysis of Energy Efficiency and Fault Tolerance. American Journal of Scholarly Research and Innovation, 13(06), 01–44. https://doi.org/10.63125/g9c9cp41
[167]. Štrumbelj, E., & Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. Knowledge and Information Systems, 41(3), 647–665. https://doi.org/10.1007/s10115-013-0679-x
[168]. Sumaiya, M. (2024). Hospital Wastewater as A Reservoir of Antimicrobial Resistance Genes: Implications for Public Health and Genomic Surveillance in The United States. Review of Applied Science and Technology, 3(04), 448–508. https://doi.org/10.63125/ahvas976
[169]. Sun, X., & Xu, W. (2014). Fast implementation of DeLong's algorithm for comparing the areas under correlated receiver operating characteristic curves. IEEE Signal Processing Letters, 21(11), 1389–1393. https://doi.org/10.1109/LSP.2014.2337313
[170]. Sundararajan, M., & Najmi, A. (2020). The many Shapley values for model explanation. In Proceedings of the 37th International Conference on Machine Learning (PMLR 119, pp. 9269–9278). https://proceedings.mlr.press/v119/sundararajan20b.html
[171]. Syed Feroz, A. (2023). Quality-Management Systems for Resilient Energy Infrastructure Fabrication. International Journal of Scientific Interdisciplinary Research, 13(10), 37–91. https://doi.org/10.63125/mbhhrm88
[172]. Syed Feroz, A. (2024). AI-Assisted NDT Analytics for Early Crack Detection in Energy-Infrastructure Pressure Vessels Using Advanced Manufacturing Data. Review of Applied Science and Technology, 14(07), 01-52.

https://doi.org/10.63125/8bdqs538

[173]. Syed Feroz, A., & Md Mahamud Pervaz, P. (2022). Root-Cause Analysis and FMEA-Driven Reliability Improvement in Gas Production Facilities: A Multi-Year Empirical Study. International Journal of Scientific Interdisciplinary Research, 12(10), 01–53. https://doi.org/10.63125/70msp878

[174]. Taha, A. A., & Malebary, S. J. (2020). An intelligent approach to credit card fraud detection using an optimized light gradient boosting machine. IEEE Access, 8, 25579–25587. https://doi.org/10.1109/ACCESS.2020.2971354

[175]. Tahmedur, R., & Rifat, C. (2023). Critical Path Analysis and Resource-Leveling Optimization for Multi-Project Engineering Portfolios. American Journal of Interdisciplinary Studies, 13(12), 01-47. https://doi.org/10.63125/8eqnea46

[176]. Tornatzky, L. G., & Fleischer, M. (1990). The processes of technological innovation. Lexington Books. https://archive.org/details/processesoftechn0000torn

[177]. Van Vlasselaer, V., Bravo, C., Caelen, O., Eliassi-Rad, T., Akoglu, L., Snoeck, M., & Baesens, B. (2015). APATE: A novel approach for automated credit card transaction fraud detection using network-based extensions. Decision Support Systems, 75, 38–48. https://doi.org/10.1016/j.dss.2015.04.013

[178]. Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harvard Journal of Law & Technology, 31(2), 841–887. https://jolt.law.harvard.edu/assets/articlePDFs/v31/Counterfactual-Explanations-without-Opening-the-Black-Box-Sandra-Wachter-et-al.pdf

[179]. Weber, M., Domeniconi, G., Chen, J., Weidele, D. K. I., Bellei, C., Robinson, T., & Leiserson, C. E. (2019). Anti-money laundering in Bitcoin: Experimenting with graph convolutional networks for financial forensics. arXiv. arXiv:1908.02591. https://arxiv.org/abs/1908.02591

[180]. West, J., & Bhattacharya, M. (2016). Intelligent financial fraud detection: A comprehensive review. Computers & Security, 57, 47–66. https://doi.org/10.1016/j.cose.2015.09.005

[181]. Whitrow, C., Hand, D. J., Juszczak, P., Weston, D., & Adams, N. M. (2009). Transaction aggregation as a strategy for credit card fraud detection. Data Mining and Knowledge Discovery, 18(1), 30–55. https://doi.org/10.1007/s10618-008-0116-z

[182]. Widmer, G., & Kubat, M. (1996). Learning in the presence of concept drift and hidden contexts. Machine Learning, 23(1), 69–101. https://doi.org/10.1007/BF00116900

[183]. Wolpert, D. H. (1992). Stacked generalization. Neural Networks, 5(2), 241–259. https://doi.org/10.1016/S0893-6080(05)80023-1

[184]. Zadrozny, B., & Elkan, C. (2002). Transforming classifier scores into accurate multiclass probability estimates. In Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 694–699). ACM. https://doi.org/10.1145/775047.775151

[185]. Zaharia, M., Xin, R. S., Wendell, P., Das, T., Armbrust, M., Dave, A., Meng, X., Rosen, J., Venkataraman, S., Franklin, M. J., Ghodsi, A., Gonzalez, J., Shenker, S., & Stoica, I. (2016). Apache Spark: A unified engine for big data processing. Communications of the ACM, 59(11), 56–65. https://doi.org/10.1145/2934664

[186]. Zhang, X., Han, Y., Xu, W., & Wang, Q. (2021). HOBA: A novel feature engineering methodology for credit card fraud detection with a deep learning architecture. Information Sciences, 557, 302–316. https://doi.org/10.1016/j.ins.2019.05.023

[187]. Zhou, Z.-H. (2012). Ensemble methods: Foundations and algorithms. Chapman and Hall/CRC. https://doi.org/10.1201/b12207

[188]. Žliobaitė, I., Pechenizkiy, M., & Gama, J. (2016). An overview of concept drift applications. In N. Japkowicz & J. Stefanowski (Eds.), Big data analysis: New algorithms for a new society (pp. 91–114). Springer. https://doi.org/10.1007/978-3-319-26989-4_4