American Journal of Interdisciplinary Studies

Volume: 4; **Issue:** 4
Pages: 210-249
eISSN: 3067-5146

AN EMPIRICAL EVALUATION OF MACHINE LEARNING TECHNIQUES FOR FINANCIAL FRAUD DETECTION IN TRANSACTION-LEVEL DATA

Mostafa Kamal[1];

[1]. VP & Unit Head, Import (RMG), City Bank PLC, Bangladesh; Email: saikatdu20@yahoo.com

Doi: 10.63125/60amyk26
Received: 19 September 2023; **Revised:** 21 October 2023; **Accepted:** 22 November 2023; **Published:** 27 December 2023

Abstract

Financial institutions increasingly rely on cloud hosted, data driven transaction monitoring, yet many fraud programs still struggle to balance fraud capture with false alert workload. This study tested how readiness determinants shape fraud detection effectiveness and compared machine learning classifiers on transaction level data in a quantitative, cross sectional, case-based design. Survey data were collected from 180 practitioners in one enterprise fraud platform case, spanning fraud and risk analysts, compliance and audit, IT and data engineering, and operations management, complemented by transaction records from the same environment. Independent variables were Data Quality, System Integration, Analytics Competency, Model Interpretability, Management Support, and Compliance Readiness; the dependent variable was Fraud Detection Effectiveness. Analysis included reliability testing, descriptive statistics, Pearson correlations, multiple regression, and an ML benchmark using precision, recall, F1, and ROC AUC with cross validation and threshold sensitivity. All constructs were reliable (Cronbach alpha .81 to .89). Fraud Detection Effectiveness was moderately high (M 3.74, SD 0.62), while System Integration was the weakest area (M 3.41, SD 0.71). Correlations were positive, strongest for Data Quality (r .62) and Model Interpretability (r .49). The regression model explained 53 percent of variance (F (6,173) 32.41, R2 .53); Data Quality (beta .36, p < .001), Model Interpretability (beta .19, p .002), Management Support (beta .17, p .004), and Analytics Competency (beta .14, p .017) were significant, while System Integration was not (beta .07, p .192). On transaction evaluation, XGBoost achieved the best balance (Precision 0.84, Recall 0.79, F1 0.81, ROC AUC 0.93) and remained stable (F1 0.80 ± 0.03). Profiling showed higher fraud rates at night (2.8 percent) and in high velocity bursts (4.6 percent). Compliance Readiness showed borderline influence (beta .09, p .052), and mid-range amounts of $120 to $500 contained 46.9 percent of fraud cases. Implications are that data governance and explainability should be treated as core controls alongside model selection, improving performance, auditability, and threshold tuning to match investigation capacity.

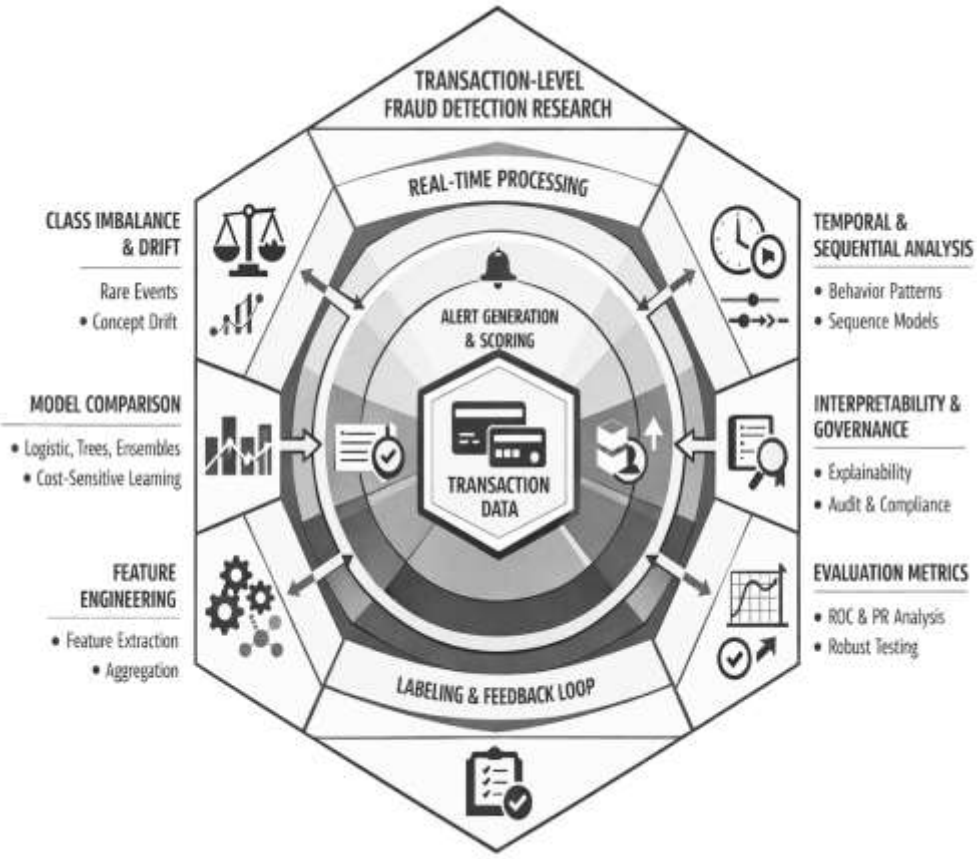
Keywords

Transaction-level fraud detection; Data quality; Model interpretability; XGBoost; Compliance readiness;

INTRODUCTION

Financial fraud can be defined as intentional deception or misrepresentation conducted to obtain an unlawful financial gain, typically by manipulating records, identities, or transaction processes within legitimate financial systems. Within digital payment ecosystems, "transaction-level fraud" specifically refers to fraudulent behavior that is observable in individual payment events (e.g., card-not-present purchases, account takeovers, synthetic identity usage) where each transaction carries attributes such as amount, merchant category, channel, timestamp, device or location signals, and authorization outcomes (Chandola et al., 2009). In practice, fraud detection is the socio-technical activity of identifying suspicious transactions with sufficient speed and accuracy to reduce monetary loss while maintaining customer experience and operational feasibility (Carcillo et al., 2018).

Figure 1: Systems Overview of Transaction-Level Fraud Detection Research



This definitional framing matters because fraud detection is not only a classification exercise but also a decision process constrained by verification latency, human investigator capacity, and asymmetric costs of errors. Peer-reviewed work has repeatedly demonstrated that fraud contexts exhibit extreme class imbalance, where legitimate transactions vastly outnumber fraudulent ones, which makes naïve accuracy metrics misleading and can conceal weak detection performance on the minority class. Research in credit card and payment fraud shows that transaction streams are also non-stationary, meaning that customer behavior and attacker tactics evolve over time, creating concept drift that can degrade model performance if monitoring and updating are not treated as core design requirements. Internationally, the significance of transaction-level fraud detection rests on the centrality of electronic payments to commerce, remittances, e-government services, and cross-border trade, where fraud losses propagate through chargebacks, compliance costs, and reputational damage, and where operational responses must be justified under audit and regulatory scrutiny (Bhattacharyya et al., 2011). In this environment, machine learning (ML) is increasingly positioned as a means to augment rule-based screening by learning complex, multivariate patterns that are not easily enumerated, while still requiring governance around explainability and accountability. Empirical evidence has also clarified that the analytical unit—single transactions versus aggregated behavioral windows—changes what

patterns become learnable, which has direct implications for how a study defines "transaction-level data" and how it constructs features for quantitative evaluation. In short, fraud detection research begins with precise definitions because the operational reality—rare events, shifting behaviors, and cost-sensitive decisions—shapes what constitutes valid evidence in empirical evaluation (Dal Pozzolo et al., 2018).

A major reason transaction-level fraud detection has become an internationally consequential research area is that payment ecosystems have expanded in volume, velocity, and heterogeneity, creating conditions where manual review alone cannot scale. The globalization of card networks and instant-payment rails increases the speed at which illicit activity can traverse jurisdictions, and the digitization of retail and service delivery increases the diversity of transaction channels and identity signals that can be exploited. In empirical studies using real-world payment data, researchers consistently observe that effective fraud detection must operate under tight latency constraints while handling streaming ingestion, near-real-time scoring, and delayed ground truth labels that arrive after investigation or chargeback processes complete. This creates a mismatch between textbook supervised learning assumptions and fraud operations, because labels are incomplete, delayed, and sometimes biased toward transactions that were reviewed rather than all transactions. Practitioner-oriented evidence reinforces that evaluation must account for class imbalance, drift, and operational constraints, rather than focusing narrowly on benchmark accuracy (Chen & Guestrin, 2016; Jinnat & Kamrul, 2021). This body of work has influenced how scholars conceptualize the "fraud detection system" as an end-to-end pipeline that includes feature generation, model scoring, alert prioritization, and feedback integration. It also explains why transaction-level detection often pairs statistical modeling with domain-driven thresholds and review policies. Methodologically, the literature shows that different algorithm families are commonly compared—logistic regression and linear models as interpretable baselines, tree ensembles as strong tabular learners, sequence models for temporal dependencies, and anomaly detection methods for rare-pattern discovery—each with trade-offs in transparency, calibration, and performance stability. In addition, imbalanced learning research highlights that resampling and cost-sensitive strategies change the effective learning objective and can improve recall on rare fraud events without collapsing precision, especially when false positives are operationally expensive. The international relevance of these findings is that payment providers in different regulatory contexts still face the same technical invariants—rarity, drift, and asymmetric costs—even when fraud typologies vary. Accordingly, robust empirical evaluation of ML techniques at the transaction level is widely treated as a prerequisite for deploying models that can be defended in audits and aligned with risk governance expectations. For this reason, an empirical, quantitative, case-study–based approach grounded in transaction-level data aligns closely with how prior research has established credible evidence in fraud analytics (Randhawa et al., 2018; Saito & Rehmsmeier, 2015).

Within the research record from 2005–2022, transaction-level fraud detection has increasingly been framed as a comparative learning problem in which multiple models must be evaluated under consistent data partitions, metrics, and cost assumptions. Early comparative studies using Decision Support Systems established that standard classifiers (e.g., logistic regression, decision trees, random forests, support vector machines) behave differently under class imbalance and feature correlation, and that performance claims require careful reporting of sensitivity, specificity, and cost-sensitive measures rather than only aggregate accuracy. Complementary work demonstrated that feature engineering choices such as transaction aggregation can materially alter detection outcomes by capturing behavioral patterns over time, which indicates that "transaction-level" evaluation must be explicit about whether the unit is a single event or an engineered representation of event sequences (Whitrow et al., 2009). As the field matured, practitioner-focused studies highlighted that data scarcity and confidentiality constraints often force researchers to use limited public datasets that may not reflect real operational distributions, which can inflate apparent performance and reduce external validity. In response, later research emphasized realistic modeling assumptions, including verification latency and evolving fraud strategies, showing that conventional batch-learning evaluations can overestimate performance in production-like settings (Venkatesh et al., 2012). At the same time, algorithmic progress in tabular learning has been shaped by ensemble methods, with gradient boosting systems offering strong predictive capability on structured features and becoming common baselines or candidate

models for fraud detection comparisons. The methodological implication supported by imbalanced learning scholarship is that resampling, threshold tuning, and cost-sensitive learning should be treated as integral parts of the modeling design, because they determine how models allocate attention to rare fraud cases. This is reinforced by empirical work in money-laundering and transaction-monitoring domains, which shows that sampling schemes and learning choices interact and can materially change detection performance and operational workload (Dal Pozzolo et al., 2014; Zulqarnain & Subrato, 2021). Additionally, evaluation science has established that metric selection must fit rare-event detection; ROC analysis is informative but can mask poor performance under extreme imbalance, motivating the use of precision-oriented metrics and careful interpretation of curves in rare-event contexts. Collectively, these studies motivate an empirical thesis structure where descriptive statistics characterize the sample, correlation analysis establishes relationships among constructs, and regression modeling tests hypotheses about factors influencing detection outcomes or organizational decision processes, while ML model comparison provides algorithmic evidence under consistent evaluation protocols (Adadi & Berrada, 2018; Akbar & Sharmin, 2022; Foysal & Subrato, 2022).

A second foundational pillar of transaction-level fraud research concerns temporal dependence and behavioral context, which has motivated the inclusion of sequential modeling and streaming architectures in fraud detection scholarship. Fraud in transaction streams is not purely a pointwise phenomenon, because attacker behavior often unfolds across multiple attempts, time windows, and merchant contexts, and legitimate customer behavior also exhibits habitual patterns that can be leveraged for discrimination. Empirical work on sequence classification for credit-card fraud detection shows that incorporating ordered transaction histories can improve detection by capturing temporal dependencies that are absent in independent-and-identically-distributed assumptions. In parallel, research on scalable streaming detection has framed fraud analytics as a near-real-time learning problem in which data pipelines, feature computation, and scoring infrastructure influence feasibility and performance, particularly at the scale of modern payment networks. These operational realities intersect directly with concept drift scholarship, which provides formal language and detection-adaptation strategies for changing distributions in streaming settings. The combination of drift, label delays, and class imbalance has been explicitly analyzed in realistic fraud modeling research, which demonstrates that a model's utility cannot be inferred only from a static test split, because performance and error profiles can shift as fraud strategies adapt. In addition, anomaly detection research offers a complementary lens, defining anomalies as observations that deviate from normal patterns, which aligns with fraud's rare-event nature while also introducing challenges around interpretability and false positives (Adadi & Berrada, 2018; Bahnsen, Aouada, et al., 2013). Isolation-based methods provide an example of anomaly scoring that can support fraud screening when labels are limited or delayed, while still requiring careful calibration to operational thresholds. Broad anomaly detection surveys further clarify that anomaly methods differ in assumptions about normality, density, distance, and isolation, and that domain constraints determine which family of methods yields reliable signals. In transaction fraud contexts, researchers have also investigated representation learning approaches such as autoencoders to compress transaction attributes into latent representations that preserve structure useful for downstream classification, reporting improvements in F1-oriented performance measures when combined with supervised classifiers (Abdul, 2023; Fawcett, 2006; Zulqarnain, 2022). Taken together, these strands justify a thesis emphasis on comparing ML techniques not only as isolated algorithms but also as modeling strategies suited to transaction sequences, streaming conditions, and operational constraints. They also support the inclusion of robustness checks as empirical evidence, because stability under distribution shift is a central requirement for trustworthy transaction-level fraud detection in real payment environments.

A third pillar concerns the statistical logic of quantitative evidence and the interpretability requirements that shape fraud decisions in regulated and audited environments (Zhang & Trubey, 2019). Transaction fraud detection produces alerts that can trigger customer friction, transaction declines, account locks, and regulatory reporting obligations, which increases the need for transparent reasoning and defensible evidence. Explainable AI research has documented that high-performing black-box models can fail to provide actionable explanations, motivating a parallel emphasis on interpretability, explanation taxonomies, and human-centered evaluation of explanations. In fraud

detection settings, this concern is not abstract; investigators and risk managers often require feature-level rationale to prioritize cases, validate patterns, and document decisions (Sun et al., 2007). Consequently, an empirical thesis that includes correlation analysis and regression modeling alongside ML comparisons aligns with the broader scientific norm of triangulating evidence: correlation matrices offer a view of association structure, regression provides hypothesis-testing logic under covariate control, and ML metrics demonstrate predictive performance under operationally relevant thresholds. Cost-sensitive and imbalanced-learning scholarship further shows that model performance must be interpreted in relation to error costs, since false positives can overwhelm review teams and degrade customer experience, while false negatives directly translate to financial loss. In empirical money-laundering detection research, the interplay between sampling and learning demonstrates that performance improvements can be artifacts of sampling choices, strengthening the argument that robustness checks should be documented rather than assumed (Ngai et al., 2011). This evidence base supports the design of quantitative instruments (e.g., Likert-scale constructs) when the study also examines organizational or operational determinants of model adoption, trust, or perceived effectiveness, because perceptions and governance practices can influence how detection systems are configured and acted upon (Han et al., 2005). In information systems scholarship, UTAUT2 provides a validated model for explaining technology acceptance and usage behavior through constructs such as performance expectancy, effort expectancy, and facilitating conditions, offering a theoretical basis for quantitatively modeling human and organizational dimensions that accompany ML deployment in practice. Empirical AML work also supports the relevance of "human-in-the-loop" and organizational process factors by discussing how ML outputs connect to investigative workflows and compliance requirements. In this combined technical and organizational framing, interpretability and statistical hypothesis testing are complementary forms of evidence: interpretability strengthens decision defensibility, while regression-based hypothesis testing strengthens causal-leaning inference within the bounds of cross-sectional, case-based quantitative research (Gama et al., 2014).

Finally, prior studies provide concrete guidance on what constitutes credible comparative evaluation at the transaction level, which directly motivates the empirical structure of this research. Comparative studies have shown that logistic regression remains a valuable baseline because it provides a transparent decision surface and interpretable coefficients, while more complex models such as tree ensembles and sequence learners can capture nonlinear interactions and temporal dependencies that improve detection performance. Ensemble methods are repeatedly documented as strong candidates in fraud contexts, with AdaBoost-based hybridization and majority voting demonstrating performance gains when carefully evaluated on benchmark and real-world datasets. Gradient boosting systems have also become central in tabular prediction tasks and are routinely used as competitive models in empirical comparisons, reinforcing the rationale for including them in model comparison sections of a fraud thesis. Streaming frameworks such as SCARFF highlight that scalability and pipeline design affect feasibility and model freshness, which is relevant when a case study seeks to reflect operational environments where transaction volumes and latency constraints are nontrivial (Misra et al., 2020). Practitioner lessons in fraud detection further argue that experimental design must respect the realities of drifting distributions, scarce labels, and confidentiality constraints, motivating careful documentation of dataset context, sampling strategy, and validation procedures within a case-study methodology (Liu et al., 2008). Relatedly, the concept drift literature offers methodological tools for describing and diagnosing shifts, strengthening the scientific justification for stability and robustness checks in empirical evaluation (Guidotti et al., 2018; He & Garcia, 2009). In addition, aggregation strategies and representation learning approaches show that feature design decisions can be as influential as algorithm choice, indicating that the study's modeling pipeline should be treated as an object of evaluation, not only the classifier family. Evaluation methodology research clarifies that ROC-oriented measures should be complemented by precision-recall–oriented analysis for rare-event detection, and that reporting should be consistent with the prevalence and operational aims of fraud screening. Under these established norms, an empirical evaluation thesis can be made more trustworthy by (a) clearly characterizing the sample and measurement reliability, (b) presenting correlation and regression evidence to test hypotheses within a quantitative design, and (c) reporting ML performance using multiple metrics that reflect rare-event decision goals and operational costs

(Jurgovsky et al., 2018).

This study is structured around a set of objectives that collectively operationalize the empirical evaluation of machine learning techniques for financial fraud detection in transaction-level data within a quantitative, cross-sectional, case-study–based design. The first objective is to clearly characterize the transaction-level fraud detection environment in the selected case context by documenting the relevant data structure, operational conditions, and respondent profile, ensuring that the empirical evidence is grounded in a well-defined setting. The second objective is to measure, using a Likert five-point scale instrument, the perceived strength of key determinants that influence fraud detection effectiveness, including data quality, system integration capability, staff analytics competency, model interpretability expectations, management support, and regulatory or compliance readiness. The third objective is to quantify the central tendencies and variability of these determinants through descriptive statistical analysis, allowing the study to establish an accurate baseline of the organizational and operational state of fraud analytics in the case environment (Jullum et al., 2020). The fourth objective is to examine the degree and direction of association among the study variables by applying correlation analysis, with particular attention to how the identified determinants relate to perceived fraud detection effectiveness and to each other within the same cross-sectional sample. The fifth objective is to test the study hypotheses through regression modeling by estimating the unique contribution of each predictor variable while controlling for overlapping effects, thereby identifying which determinants significantly explain variance in fraud detection effectiveness in the case context. The sixth objective is to conduct an empirical comparison of selected machine learning techniques for fraud detection using transaction-level data, reporting performance through fraud-appropriate metrics such as precision, recall, F1-score, and ROC-AUC so that algorithmic effectiveness can be assessed in a consistent and transparent manner. The seventh objective is to strengthen the trustworthiness of the empirical findings by presenting fraud-pattern profiling and risk signature results that describe how fraudulent and legitimate transactions differ across meaningful behavioral and transactional dimensions in the case dataset. The eighth objective is to validate the consistency of model outcomes through robustness and stability checks, including performance variability across multiple validation splits and sensitivity to threshold settings, ensuring that the reported model comparisons reflect dependable behavior rather than isolated results. The ninth objective is to provide decision-logic evidence by reporting explainability results, such as influential feature patterns and model reasoning indicators, to support interpretability and practical auditability within fraud detection decision processes. The final objective is to synthesize findings across statistical hypothesis testing and machine learning evaluation by mapping results directly back to the research questions and objectives, maintaining alignment between the study design, the empirical analyses, and the measurable outcomes derived from the case-study setting.

LITERATURE REVIEW

Financial fraud detection has become a central research domain within data mining, information systems, and financial risk management because modern payment ecosystems generate massive volumes of transaction-level data where fraudulent behavior is rare, adaptive, and operationally costly to miss. The literature frames fraud detection as a high-stakes classification and decision-support problem in which models must separate legitimate from illegitimate transactions under conditions of extreme class imbalance, heterogeneous feature types, and shifting behavioral patterns across customers, channels, and merchant contexts. Transaction-level data typically include numerical, categorical, temporal, and contextual signals, and prior scholarship emphasizes that the value of such data depends heavily on preprocessing, feature construction, and the alignment of analytical objectives with operational realities such as alert handling capacity, verification latency, and compliance documentation. As research matured, studies increasingly compared traditional statistical models and classical machine learning algorithms with more advanced ensemble and deep learning approaches, showing that algorithm choice alone does not determine success; rather, performance is shaped by sampling strategies, threshold selection, cost-sensitive learning, and the stability of models under real-world distribution changes. Accordingly, evaluation practices in the fraud literature stress the limitations of accuracy and highlight the need for metrics that reflect rare-event detection quality, such as precision, recall, F1-score, and AUC measures, combined with analyses that account for false-

positive workload and false-negative loss exposure. In parallel, a growing stream of work examines explainability and governance in fraud analytics, noting that interpretable decision logic and auditability are essential when automated decisions affect customers and trigger regulated processes, which has motivated the inclusion of model transparency techniques and human-in-the-loop considerations in empirical studies. Alongside technical contributions, the literature also recognizes that fraud detection systems are implemented within organizational environments, where data quality management, system integration, staff competency, and managerial support influence how models are deployed, trusted, and acted upon, creating a need for empirical research designs that integrate technical performance evidence with measurable organizational determinants. Within this context, an objective-aligned literature review must synthesize findings across fraud typologies, transaction data characteristics, algorithmic families, evaluation standards, interpretability requirements, and theoretical perspectives explaining technology effectiveness and adoption in financial institutions, so that the present study can ground its hypotheses and empirical evaluation strategy in established scholarly knowledge while maintaining alignment with its quantitative, cross-sectional, case-study–based methodology.

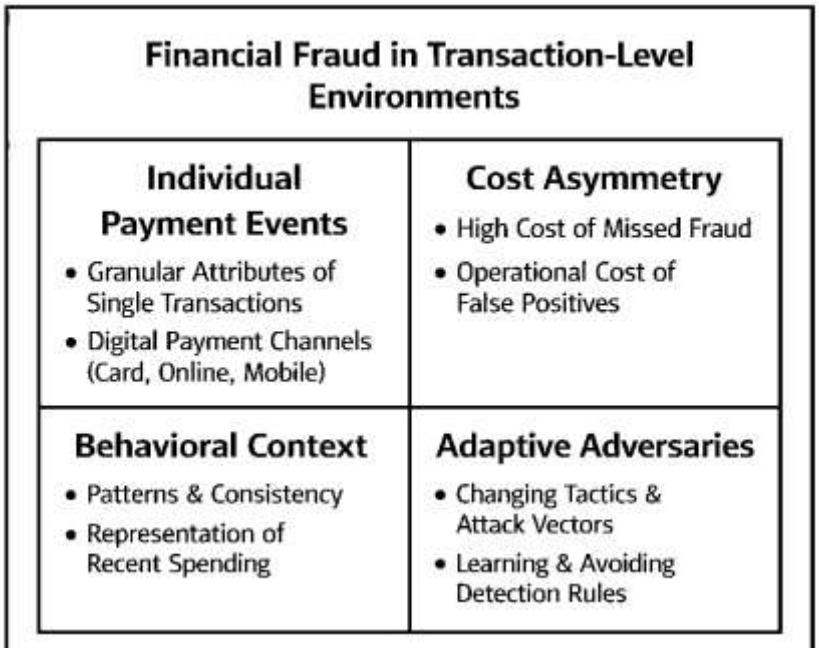
Financial Fraud in Transaction-Level Environments

Financial fraud in transaction-level environments refers to deliberate attempts to obtain unauthorized value through individual payment events that move across digital rails such as card payments, online banking transfers, mobile wallets, and merchant acquiring networks. At this level of granularity, fraud is expressed through the attributes and context of each transaction—amount, timestamp, channel, merchant type, location and device signals, authentication outcome, and behavioral consistency with the account's prior spending profile. The literature characterizes this domain as operationally demanding because detection is expected to occur under tight time constraints while preserving legitimate customer activity and minimizing friction (Hammad & Mohiul, 2023; Hasan & Waladur, 2023). A central feature of transaction-level fraud is the strong asymmetry of error costs: a missed fraud event translates directly into financial loss, while an incorrectly flagged legitimate transaction generates customer dissatisfaction, manual review workload, and potential revenue interruption. For this reason, scholars have argued that fraud detection should be treated as a decision problem rather than a purely predictive exercise, because the "best" model depends on how institutions value losses, investigation resources, and customer experience. This framing becomes more prominent in environments where millions of transactions must be screened continuously, requiring systems that can prioritize alerts and support rapid intervention. In addition, transaction-level fraud is shaped by adversarial adaptation: fraudsters alter tactics to exploit new channels, avoid detection triggers, and replicate legitimate behavioral patterns, which intensifies the need for methods that learn from patterns at scale and support practical monitoring of false alarm rates. These realities are frequently described through performance criteria that explicitly recognize the trade-off between fraud coverage and operational cost, highlighting that evaluation must be aligned with the economics of detection and response in real payment operations (Hand et al., 2007).

Transaction-level fraud is also methodologically distinctive because its data structure is inherently imbalanced, heterogeneous, and behavior-dependent. Fraud cases are rare relative to legitimate activity, and the signals that distinguish them often emerge only when single events are interpreted in the context of an account's recent history. This has motivated the common practice of representing transaction behavior through aggregation and sequencing, where raw transaction attributes are combined with short-window summaries such as transaction counts, total spend, merchant diversity, and velocity indicators. Research emphasizing aggregation has shown that a single transaction can be insufficient to identify fraud reliably because fraud cues frequently involve deviations from typical spending routines rather than isolated attribute values. Consequently, transaction-level environments often require engineered features that encode behavioral consistency and temporal proximity, allowing models to compare new activity against an account's recent baseline. Empirical studies have demonstrated that transaction aggregation strategies can materially improve detection by capturing patterns that are invisible in raw fields alone, such as rapid bursts of spending, unusual merchant grouping, or abrupt changes in spend intensity. From an organizational standpoint, these engineered representations also support communication between analytics teams and fraud investigators, because

aggregated features can be linked to intuitive risk narratives (e.g., "unusual velocity in a short period"). The literature further indicates that model selection is inseparable from representation design: certain methods may appear to perform well on raw data but lose advantage once behavior-oriented features are introduced, or vice versa. This evidence supports the view that transaction-level fraud detection should evaluate not only algorithms but also the feature engineering strategies that reflect how fraud manifests in operational payment streams (Jha et al., 2012).

Figure 2: Financial Fraud In Transaction-Level Environments



Beyond representation, transaction-level fraud environments raise persistent concerns about deployment practicality, stability, and real-world performance generalization. Because payment systems operate continuously and fraud behavior shifts, models are expected to remain effective under changing distributions, evolving customer habits, and varying channel risk exposure. As a result, many transaction-level studies compare hybrid approaches that combine machine learning classification with rules, thresholds, and workflow constraints to achieve usable alert volumes and consistent decision quality. Earlier work proposed hybrid detection models that integrate multiple learning components to balance sensitivity and specificity under operational conditions, indicating that single-model solutions often struggle to satisfy both loss reduction and manageable false-positive workloads (Krivko, 2010; Rifat & Rebeka, 2023; Zulqarnain & Subrato, 2023). More recent comparative studies using established fraud datasets and practical feature windowing have reinforced that model effectiveness depends on validation design, threshold tuning, and feedback handling, especially when data arrives as a stream rather than as a static batch (Dornadula & Geetha, 2020). At the system level, survey research has consolidated these lessons by documenting common fraud types, the practical limitations of purely rule-based systems, and the technical challenges that emerge from real-time requirements and evolving attacker strategies, reinforcing the need for comprehensive evaluation that aligns technical metrics with operational feasibility (Abdallah et al., 2016). Taken together, the literature positions transaction-level fraud detection as a socio-technical domain where trustworthy empirical evaluation must account for cost asymmetry, behavioral representation, and operational stability, establishing a clear foundation for studies that compare machine learning techniques within bounded case environments while emphasizing rigorous, transaction-centered evidence.

Hazard Communication Mechanisms and Worker Comprehension in Industrial Settings

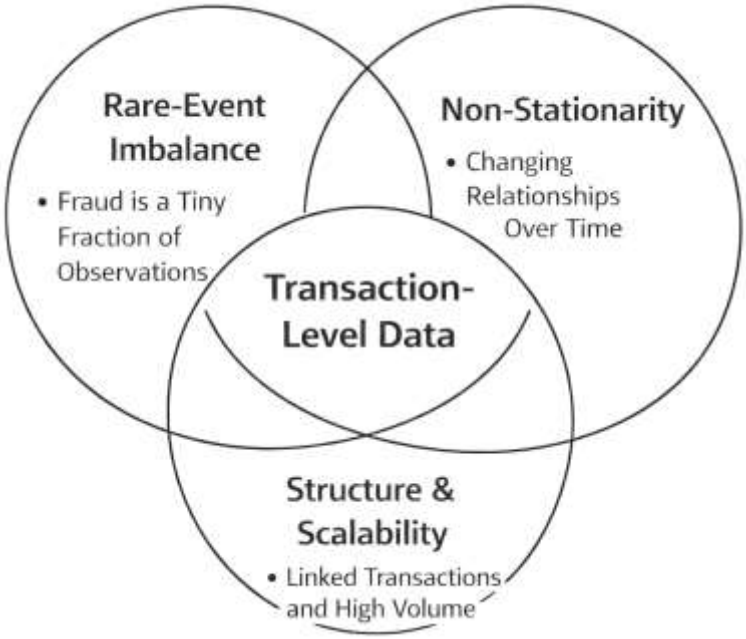
Transaction-level datasets used for financial fraud detection are defined by high volume, high velocity, and high heterogeneity, where each record captures a single payment event while the meaning of that event depends heavily on context. Typical fields combine numeric attributes (amount, balance, limits),

categorical markers (merchant category, channel, authorization route), temporal indicators (time-of-day, day-of-week, inter-transaction time), and device or geo-proxies, producing mixed data types that require careful encoding and preprocessing before modeling. The literature consistently treats this setting as "rare-event learning," because fraudulent outcomes may represent far below one percent of observations, which makes naive learning objectives misleading and encourages models to optimize toward the majority class. In practice, this imbalance interacts with operational realities: a classifier that is highly "accurate" in aggregate can still be unusable if it produces either excessive false alarms or misses fraud clusters that matter financially. Transaction logs also include redundancies and repeated behavioral patterns (routine purchases, recurring bills), meaning that raw fields alone often fail to represent the behavioral deviations that investigators actually associate with suspicious activity. For this reason, transaction-level fraud detection research emphasizes representation design as a foundational challenge, not a peripheral step, because feature choices determine whether models can see changes in velocity, periodicity, and merchant-consumption structure. A well-established stream of work shows that extended feature engineering—especially aggregation across short windows and the creation of periodic/behavioral indicators—can substantially improve detection outcomes by converting isolated transactions into behavior-aware signals, which is more consistent with how fraud manifests in real payment behavior (Bahnsen et al., 2016). This makes transaction-level data a domain where methodological rigor begins with data characterization, because the statistical properties of the dataset shape what "good performance" can even mean.

A second defining challenge is non-stationarity: transaction-level behavior evolves due to seasonality, customer lifestyle changes, merchant ecosystem dynamics, and adversarial adaptation by fraudsters. This phenomenon is commonly conceptualized as concept drift, where the statistical relationship between features and fraud labels changes over time, making static models degrade even if they were strong at deployment. Transaction streams also present label-timing complications that are far less prominent in many textbook classification problems. In real fraud operations, only a small subset of alerts are reviewed promptly by investigators, while many labels become available later through customer disputes or chargeback processes, creating verification latency that can distort what the model "learns" if feedback and delayed labels are treated as equivalent. The practical implication is that the dataset is not merely imbalanced; it is also partially observed in time, and the sample that receives immediate labels is not random but shaped by the system's own alerting policy. Research addressing these realities highlights that learning strategies must account for drift and delayed supervision simultaneously, because the data-generation process is coupled with the detection workflow rather than being an independent labeling pipeline (Dal Pozzolo et al., 2015). In addition, contemporary transaction environments increasingly require incremental or online adaptation to maintain effectiveness, especially when payment channels expand (e-commerce, mobile, tokenized transactions) and fraud tactics mutate quickly. Methods that combine window-based updating, resampling, and cost-aware learning are therefore positioned as responses to intrinsic transaction-data properties (imbalance, drift, borderline cases, and noise), rather than optional "enhancements" (Somasundaram & Reddy, 2019). In short, transaction-level fraud detection is a moving-target inference task embedded in organizational response constraints, which makes stability under time change an essential data-driven concern.

A third set of transaction-level challenges relates to structure and scalability: transactions are not independent events in practice, because they are linked through accounts, merchants, devices, and interaction histories that can be modeled as relational networks. This relational dimension matters because fraud can emerge as coordinated behavior (shared devices, connected merchant rings, repeated attack pathways), where suspiciousness is expressed through connectivity patterns rather than isolated attribute thresholds. Consequently, fraud detection data are often better understood as a socio-technical trace of interactions in a payments ecosystem, motivating approaches that enrich tabular records with network-derived signals and neighborhood behavior indicators. Research demonstrates that network-based extensions can add discriminative information by capturing relationships among entities that conventional feature sets overlook, particularly when fraud is organized or distributed (Van Vlasselaer et al., 2015).

Figure 3: Transaction-Level Data Challenges In Financial Fraud Detection



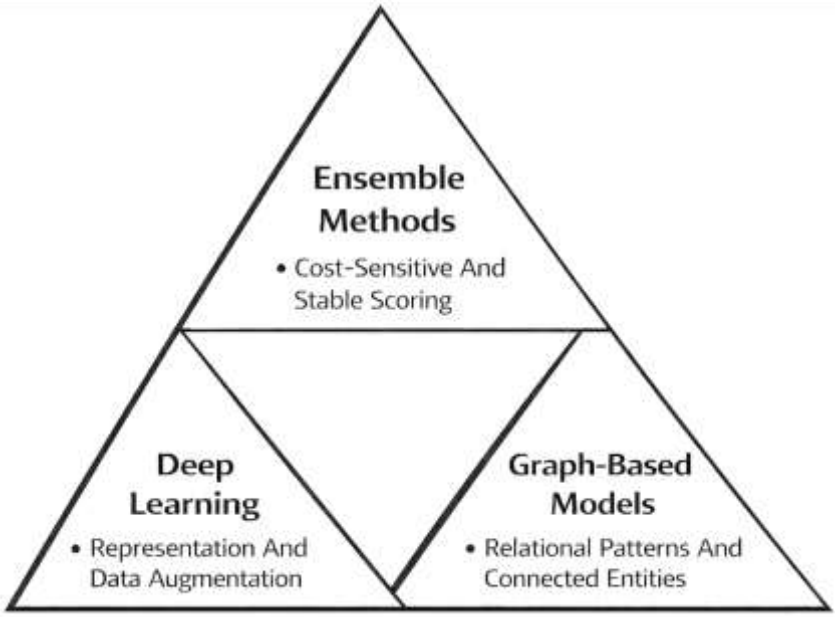
At the same time, the operational scale of transaction streams forces engineering trade-offs: models must produce decisions in near-real time, integrate with streaming infrastructure, and remain performant under massive throughput, which pushes researchers to evaluate end-to-end frameworks rather than only algorithms. This is where scalability and real-time readiness become data characteristics in their own right, because the "shape" of the data (continuous stream, large volume, rapid arrival) constrains feasible training and scoring pipelines. Framework-oriented studies show that effective streaming fraud detection must jointly handle imbalance, non-stationarity, and feedback latency while remaining computationally scalable, reinforcing that transaction-level data challenges are both statistical and infrastructural (Carcillo et al., 2018). Together, these streams of evidence justify empirical evaluations that treat transaction-level fraud detection as an integrated problem of representation, temporal validity, and operational feasibility, aligning directly with the present study's emphasis on trustworthy comparative modeling and robust evidence.

Machine Learning Techniques for Transaction-Level Financial Fraud Detection

Transaction-level fraud detection has matured into a model-comparison problem in which multiple supervised learners are evaluated under severe class imbalance, shifting behavior, and asymmetric misclassification costs. A common technical baseline is the family of tree-based ensembles and margin-based classifiers that can exploit heterogeneous feature sets (amount, time, merchant category, device proxies, velocity variables, and engineered interaction terms) while remaining robust to noisy attributes. Within this stream, cost-sensitive learning is frequently emphasized because operational loss is not symmetric: a false negative may permit direct monetary leakage, while a false positive may create customer friction, manual review workload, and reputational cost. A representative technique-level contribution is the two-stage, cost-sensitive pipeline in which similarity-based behavioral matching is conducted first and a dynamic random forest is applied afterwards to address evolving cardholder patterns; the explicit objective is to increase fraud "damage prevention" rather than inflate accuracy on an imbalanced dataset (Nami & Shajari, 2018). This family of approaches reframes model selection around business-aligned targets by embedding cost functions, minimum-risk rules, or class-weighted optimization directly into training and thresholding. In practice-oriented evaluations, these models are commonly contrasted with logistic regression and other generalized linear baselines because the interpretability of coefficients and the simplicity of calibration remain valuable for governance. Even when more complex learners outperform in recall-oriented metrics, the strongest technique narratives in the literature highlight that the winning model is often the one that provides stable performance across time windows and maintains controllable false-positive rates at operationally feasible alert

volumes. Under this perspective, "best" technique is not solely the highest AUC; it is the method whose learning objective and decision threshold can be aligned with how institutions absorb risk, investigate cases, and manage customer experience (Fiore et al., 2019).

Figure 4: Machine Learning Technique Clusters For Transaction-Level Financial Fraud Detection



Deep learning techniques expand the modeling space by shifting emphasis from hand-crafted predictors to representation learning, particularly when fraud signatures are subtle, non-linear, and distributed across many weak signals. Autoencoder-style designs and sequence-aware models are regularly motivated by the need to capture latent structure in normal transactions so that deviations can be surfaced as anomalies or fed into downstream classifiers. A second deep-learning direction treats the data imbalance problem as a generative modeling challenge: rather than relying only on resampling heuristics, a generator is trained to synthesize plausible minority-class examples that enrich the decision boundary of the discriminator. In credit-card fraud contexts, generative adversarial networks (GANs) have been proposed to produce realistic fraudulent transaction patterns so that conventional classifiers trained on the augmented data can improve discrimination, particularly in regimes where fraud examples are scarce and diverse (Liu et al., 2020). The technical logic here is that a better approximation of the minority manifold can reduce overfitting to a few observed fraud modes and can increase recall without collapsing precision. At the same time, the technique literature stresses that generative enrichment is not purely a data trick; it is a modeling choice that must be validated against the risk of generating artifacts that are statistically plausible but operationally implausible. For fraud detection, this concern is amplified because fraud is adversarial and strategic: synthetic instances must preserve constraints of transaction systems (limits, channel rules, category codes) and must remain consistent with cross-field dependencies that investigators expect. Consequently, deep-learning technique discussions often position GAN augmentation as an "assistive" module that improves traditional learners and ensemble stacks, rather than a replacement for governed scoring pipelines. In this sense, representation learning and generative modeling broaden the set of candidate techniques while keeping the evaluation logic anchored to threshold stability, error-cost tradeoffs, and defensible decision behavior under real transaction constraints (Fiore et al., 2019).

A third technique cluster models transaction streams as networks, reflecting that fraud rarely exists as isolated points; it often manifests through relational patterns (shared devices, merchant rings, mule accounts, coordinated bursts, and repeated interactions). Graph-based anomaly detection and graph neural networks (GNNs) have therefore become prominent because they encode dependencies that tabular models can miss, especially when the suspicious signal is "who transacts with whom" rather than "what a single transaction looks like." A broad synthesis of graph-based anomaly detection for

fraud emphasizes that connectivity structure, community behavior, and link patterns provide complementary evidence to intrinsic transaction features, while also introducing new challenges such as scale, dynamic graphs, and the need for domain-grounded graph construction choices (Pourhabibi et al., 2020). At the method level, GNN-based fraud detectors can be undermined by an "inconsistency" issue: neighbors in a transaction graph are not always homophilous, and malicious actors may camouflage relations or connect to legitimate nodes; technique adaptations therefore include neighbor filtering, relation-aware attention, and context-aware embedding designs (Liu et al., 2020). In parallel, probabilistic sequential decision techniques highlight that fraud detection can be framed as quickest change detection in a monitored purchasing process, where the goal is to trigger an alarm soon after a latent "fraud time" while controlling false alarms through optimal stopping and personalized thresholds (Buonaguidi et al., 2022). Together, these relational and sequential approaches strengthen the methodological foundation for transaction-level fraud detection by treating fraud as a pattern unfolding over entities and time, rather than a static classification label. They also connect naturally to rigorous evaluation designs because they invite stability checks across time segments, scenario testing under behavior shifts, and decision-rule inspection at the alert threshold—capabilities that can complement your thesis's regression-based hypothesis testing and ML model comparison in a case-study setting (Pourhabibi et al., 2020).

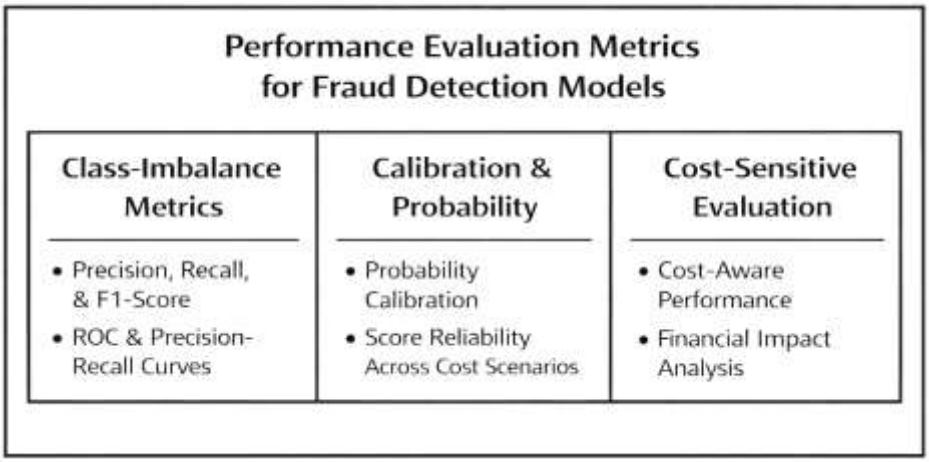
Performance Evaluation Metrics for Fraud Detection Models

Selecting appropriate evaluation metrics is a central methodological requirement in transaction-level fraud detection because the class distribution is typically highly skewed and the operational meaning of errors is asymmetric. The literature emphasizes that metrics such as overall accuracy can be misleading in rare-event contexts, since a model can achieve very high accuracy by predicting the majority "legitimate" class while failing to identify fraudulent transactions at a useful rate. For this reason, fraud studies commonly report precision, recall, and the F1-score, which quantify different aspects of minority-class performance and directly connect to operational realities such as the proportion of alerts that are truly suspicious and the proportion of fraud successfully captured. Precision reflects the quality of alerts delivered to investigators, while recall reflects coverage of fraudulent activity, and their balance is often summarized by F1 when a single number is required. In addition, curve-based metrics are routinely used to evaluate performance across different decision thresholds, especially when institutions tune thresholds to match investigation capacity or risk appetite. A key methodological contribution in this area is the formal relationship between ROC space and precision–recall (PR) space, demonstrating that the same classifier can appear strong under ROC analysis while offering a less favorable picture in PR space when positive cases are rare, which is common in fraud detection; this motivates the frequent use of PR curves and area-under-PR summaries for highly imbalanced settings (Davis & Goadrich, 2006). The same perspective supports threshold-aware reporting, because fraud operations typically require a concrete threshold that determines an alert queue, rather than a purely rank-based comparison. Accordingly, studies increasingly treat metric choice as a design decision that must match the operational objective, meaning that the evaluation section must justify why the chosen metrics are appropriate for rare-event screening and how they map onto actionable decision rules in transaction monitoring workflows.

Beyond discrimination, fraud detection evaluation increasingly incorporates probability quality because many operational decisions depend on calibrated risk scores rather than hard labels. In practice, institutions may apply different thresholds for different channels, customers, merchants, or transaction amounts, which requires that predicted probabilities represent reliable estimates of risk. The literature warns that probability estimates produced by supervised models can be systematically distorted under extreme imbalance, even when classification performance seems acceptable, creating a gap between "good classification" and "useful decision support." In a prominent imbalanced-learning analysis, researchers show that class probability estimates can be unreliable for minority instances and that standard imbalance-handling methods used for classification do not automatically correct calibration; this reinforces the need for probability-focused evaluation using calibration-aware scoring rules and class-conditional checks (Wallace & Dahabreh, 2012). To address this, calibration methods such as beta calibration have been proposed as practical, well-founded procedures to improve probability estimates across diverse classifier families, strengthening the reliability of score-based

decisions and enabling more consistent threshold selection under varying cost assumptions (Kull et al., 2017).

Figure 5: Performance Evaluation Metrics For Transaction-Level Fraud Detection Models



In fraud detection contexts, this matters because calibration affects downstream processes such as alert prioritization, triage rules, and analyst workload planning; a poorly calibrated model can generate unstable alert volumes when base rates shift, even if its ranking ability is strong. As a result, modern evaluation practice often combines discrimination metrics (e.g., PR-AUC, ROC-AUC) with probability diagnostics (e.g., calibration behavior by score bins), allowing researchers to argue that the model not only separates fraud from non-fraud but also supports dependable operational decision-making. This dual emphasis increases the credibility of model comparisons because it reduces the risk that the "best" model is chosen on the basis of a metric that is insensitive to practical deployment behavior.

A further requirement for fraud evaluation is cost alignment, since the financial impact of errors is rarely uniform: the cost of missing a high-value fraudulent transaction differs from missing a low-value one, and the cost of false positives accumulates through investigation labor, customer friction, and potential revenue disruption. This motivates cost-sensitive evaluation frameworks that convert model outcomes into monetary terms or savings-style measures, enabling model selection that better reflects fraud program objectives. One influential approach explicitly incorporates real fraud costs through Bayes minimum risk, proposing a comparison measure that represents monetary gains and losses rather than only statistical accuracy, which supports the selection of models that minimize expected financial harm under realistic fraud conditions (Bahnsen, Stojanovic, et al., 2013). Similarly, cost-sensitive ensemble frameworks evaluate models through cost-oriented outcomes, emphasizing that the same classifier can be "better" or "worse" depending on how costs are specified and how thresholds are chosen to control false positives (Olowookere & Adewale, 2020). In transaction-level fraud detection, this cost perspective is essential because model performance must be judged not only by how well it identifies fraud but also by whether it produces an alert stream that can be processed and that yields net savings after operational expenses. Consequently, credible evaluation practice increasingly includes thresholding strategies, scenario-based comparisons (e.g., high-recall versus low-false-positive operating points), and sensitivity analysis on cost assumptions. When combined with PR-based discrimination reporting and calibration evidence, cost-sensitive evaluation strengthens trustworthiness by demonstrating that the model comparison aligns with the realities of financial loss prevention and organizational capacity, rather than optimizing an abstract metric that does not reflect how fraud detection systems function in real transaction environments.

Theoretical and Conceptual Foundations for Empirical Fraud-Detection Evaluation

A rigorous empirical study of transaction-level fraud detection benefits from a theory-informed lens that explains *why* organizations adopt (and trust) machine-learning (ML) detection systems and *how* those systems translate into measurable performance outcomes. In this thesis, an organizational adoption perspective helps justify constructs such as data readiness, infrastructure adequacy, analytics capability, and governance maturity as antecedents to effective fraud detection. Prior adoption research shows that organizational decisions around advanced digital solutions are shaped by perceived

technological attributes (e.g., relative advantage, complexity, security/trust), internal organizational conditions (skills, resources, managerial support), and contextual pressures that define acceptable risk-taking and compliance boundaries. Work reconceptualizing adoption drivers for cloud computing, for example, emphasizes the operationalization of innovation characteristics (compatibility, relative advantage, complexity, security & trust) as determinants that shape an organization's willingness to implement data-intensive systems (Stieninger et al., 2014). This framing is highly relevant to fraud analytics because ML detectors are similarly infrastructure-dependent and trust-sensitive: they require reliable pipelines, scalable compute, secure data handling, and stable integration with transaction authorization workflows. In empirical terms, this theory support helps justify hypotheses that link "technological readiness" and "organizational capability" to dependent variables such as *fraud-detection effectiveness* and *investigation efficiency*. A practical way to operationalize the theory in a quantitative model is to treat system effectiveness as an outcome of measurable predictors (survey constructs and operational indicators), for instance:

$$\text{FDE} = \beta_0 + \beta_1(\text{TechReadiness}) + \beta_2(\text{DataQuality}) + \beta_3(\text{SkillCapability}) + \beta_4(\text{Governance}) + \varepsilon$$

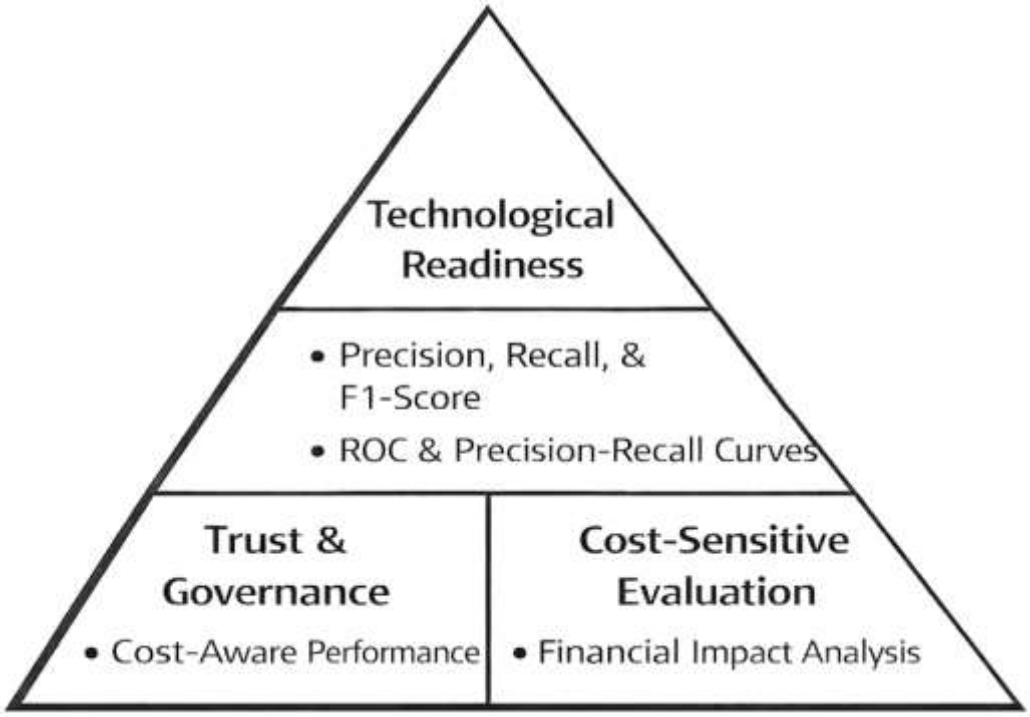
where FDE denotes fraud-detection effectiveness (e.g., perceived/observed improvement in capture rate, reduced false alerts, faster decision cycles). This equation aligns the thesis design (Likert-scale measurement, correlation, regression) with a defensible theoretical explanation for why the predictors matter in transaction-level fraud contexts.

Beyond general adoption logic, data-analytics readiness models provide a direct bridge between organizational conditions and the *quality of the evidence* produced by ML fraud systems. Transaction-level fraud detection depends not only on algorithm choice, but also on the organization's ability to assemble timely data, engineer behavioral features, maintain feedback loops, and manage model drift under evolving fraud tactics. Research using readiness-oriented frameworks grounded in technology–organization–environment thinking highlights that "readiness" is not a single attribute; it is a configuration of resource availability, infrastructure, human skills, and managerial commitment that determines whether data-driven initiatives can deliver reliable results at scale (Alazzam et al., 2021). In fraud detection, readiness becomes a credibility factor because weak data pipelines or limited analytic expertise can produce unstable models whose performance cannot be replicated across time windows or customer segments. Similarly, empirical adoption studies of big data analytics in operational domains show that adoption intention is shaped by distinct technological, organizational, and environmental drivers—and that these drivers can be tested quantitatively using survey instruments and structural relationships (Lai et al., 2018). Translating this into the present study, the conceptual logic is that ML model performance metrics (precision/recall, AUC, cost savings) are downstream expressions of upstream readiness and adoption conditions; therefore, the thesis can defend why it measures both (a) organizational determinants through Likert constructs and (b) algorithmic outputs through comparative metrics tables. A complementary "model-side" formulation is logistic regression for the probability that a transaction is fraudulent:

$$P(Y = 1 \mid X) = \frac{1}{1 + e^{-(\beta_0 + \beta^{\top} X)}}$$

where Xrepresents transaction features (amount, velocity, merchant/channel codes, engineered behavior indicators). This formula supports the study's regression component while remaining consistent with transaction-level risk scoring used in practice.

Figure 6: Theoretical And Conceptual Foundations For Empirical Fraud-Detection Evaluation



In Addition, trust and governance perspectives strengthen the thesis by explaining why *explainability, institutional pressures, and accountability* affect real-world acceptance of ML fraud decisions. Fraud detection is a high-stakes, adversarial domain in which analysts, compliance teams, and customer-facing units must rely on model outputs to block or approve financial activity; thus, the "human trust" dimension becomes central to whether ML recommendations are used consistently and whether decision rules remain stable under uncertainty. Empirical reviews on trust in AI show that organizational uptake depends heavily on perceived reliability, transparency, and the alignment between the system's behavior and users' expectations, making trust a measurable factor that can influence adoption and sustained use of AI-enabled decision systems (Glikson & Woolley, 2020). At the same time, organizations face external pressures—regulatory expectations, industry norms, customer demands—that can accelerate or constrain adoption of analytics-powered AI. Evidence from institutional-theory research demonstrates that coercive, normative, and mimetic pressures shape how firms mobilize resources and skills to adopt analytics-powered AI capabilities, reinforcing the idea that "environmental context" can be modeled as a predictor of implementation success and performance (Bag et al., 2020). For fraud detection, these pressures translate into requirements for auditability, defensible thresholds, consistent treatment of customers, and documented control logic. A cost-aligned evaluation expression further formalizes why trust and governance matter:

$$\text{ExpectedLoss} = C_{FN} \cdot FN + C_{FP} \cdot FP$$

where $FN$are missed fraud cases and $FP$are false alerts, and $C_{FN}$and C_{FP}represent institution-specific costs. Because governance determines how thresholds are chosen and how errors are tolerated, this formulation supports hypotheses linking governance maturity and explainability practices to measurable reductions in loss and operational burden. Together, these theory-backed constructs make the empirical evaluation more trustworthy because they connect model performance to organizational readiness, trust, and environmental accountability rather than treating fraud detection as an algorithm-only exercise.

Integrated Conceptual Framework and Research Gaps

A defensible empirical framework for transaction-level fraud detection must explain effectiveness as more than an algorithmic score, because real detection systems are socio-technical artifacts that transform raw transaction traces into risk decisions under operational constraints. A useful theoretical

anchor is the information-systems success (IS-success) tradition, which treats net benefits as the downstream result of system quality and information quality, mediated by use, satisfaction, and service support. In fraud analytics, "system quality" translates into pipeline reliability, latency, integration with authorization and case-management tools, and stability under high throughput, while "information quality" reflects the accuracy, completeness, timeliness, and consistency of transaction attributes, labels, and engineered features. Evidence from data-warehousing research shows that information and system quality are not abstract ideals; they have identifiable antecedents (e.g., accuracy, reliability, accessibility) that explain substantial variance in perceived quality and, by extension, downstream outcomes in analytical settings (Nelson et al., 2005). Similarly, the IS-success literature synthesizes how different success dimensions relate and how measurement choices shape empirical conclusions, which is crucial when a fraud study combines operational metrics (precision/recall) with survey-based constructs (e.g., perceived effectiveness, trust, readiness) (Petter et al., 2008). A compact way to formalize this in a cross-sectional thesis is to define fraud-detection effectiveness (FDE) as a latent or composite outcome influenced by measured quality and capability predictors, estimated through regression and supported by descriptive and correlational evidence. For example, the empirical backbone can be represented as:

$$\text{FDE} = \beta_0 + \beta_1(\text{InfoQ}) + \beta_2(\text{SysQ}) + \beta_3(\text{AnalyticCap}) + \beta_4(\text{Governance}) + \varepsilon$$

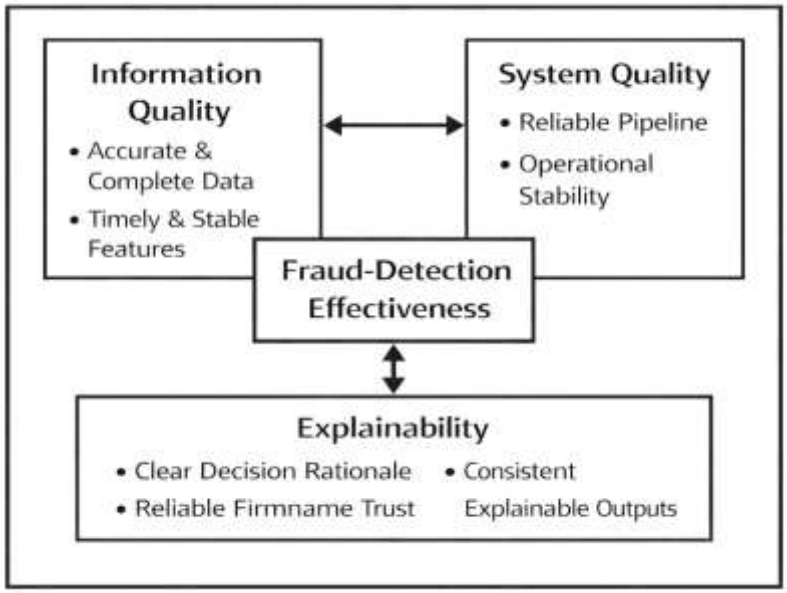
where InfoQ and SysQ can be measured via Likert items aligned to IS-success measurement practice, and AnalyticCap and Governance capture organizational conditions that determine whether models are used consistently. This framing increases construct clarity because it ties "effectiveness" to measurable system and information properties rather than presenting model performance as self-explanatory evidence (Ribeiro et al., 2016).

A second foundation for the conceptual framework is the recognition that modern fraud detection requires explainability and accountability, not only predictive power. When detection models operate as black boxes, stakeholders may distrust alerts, override recommendations, or hesitate to operationalize automated actions (e.g., declines, step-up authentication), especially when decisions must be documented for audit review. Responsible-AI research consolidates the argument that explainability is audience-dependent and must be considered alongside other principles such as accountability and transparency, which are particularly salient in financial decision-making (Barredo Arrieta et al., 2020). In practical terms, this suggests treating explainability not merely as a post-hoc visualization, but as a measurable dimension of decision quality that affects adoption, consistent use, and defensible operations. The conceptual link can be formalized by defining "decision-logic evidence" as an enabling condition that strengthens the relationship between model scores and operational action. One operational mechanism is local explanation: a transaction-level explanation highlights which features drove a risk score for a specific alert, supporting investigator triage and managerial oversight. Work on local surrogate explanations provides a concrete methodological basis for producing human-interpretable rationales around individual predictions, enabling a fraud system to present both a risk score and an explanation artifact (Ribeiro et al., 2016). In a conceptual framework, explainability can therefore act as a mediator or moderator between technical performance and perceived usefulness/trust, particularly in a case study where investigators and compliance stakeholders evaluate whether model outputs "make sense." A simple moderation form (estimable via regression) is:

$$\text{FDE} = \beta_0 + \beta_1(\text{ModelPerf}) + \beta_2(\text{XAI}) + \beta_3(\text{ModelPerf} \times \text{XAI}) + \varepsilon$$

where ModelPerf can be represented by the best-performing model's PR-AUC or F1, XAI is an explainability construct measured via Likert items (clarity, auditability, actionability), and the interaction term tests whether explainability strengthens how performance translates into credible effectiveness. This is consistent with the idea that "good scores" alone do not ensure value unless the organization can understand and act on decisions reliably (Barredo Arrieta et al., 2020).

Figure 7: Integrated Conceptual Framework For Transaction-Level Fraud Detection



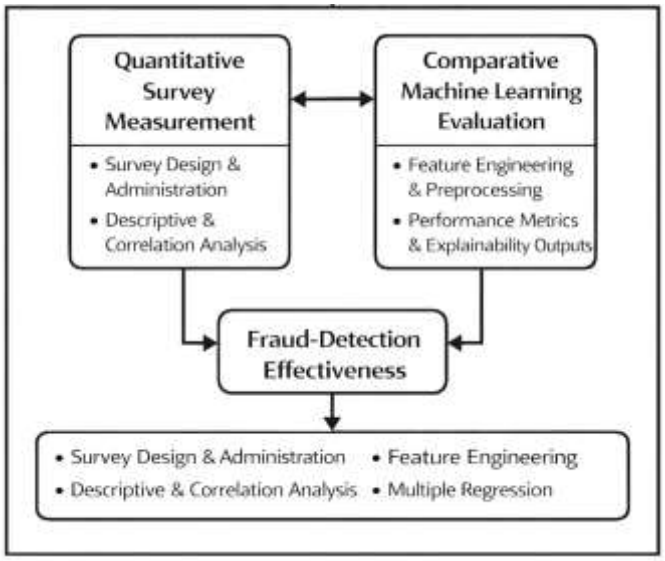
Within this combined lens, several research gaps become visible and motivate the present thesis structure. First, many fraud studies emphasize algorithmic comparisons while under-specifying how information quality and system quality shape results, which can lead to fragile conclusions when datasets differ in labeling practices, missingness, or feature stability; integrating explicit InfoQ and SysQ constructs responds to this gap by treating data and pipeline conditions as measurable determinants of effectiveness rather than hidden assumptions (Jeyaraj, 2020). Second, evaluation research frequently reports a single "best" classifier without connecting model performance to organizational uptake and net benefits; a conceptual framework grounded in IS-success clarifies that realized benefits depend on use, workflow integration, and satisfaction, and it encourages mapping results back to operational outcomes and stakeholder acceptance. Third, the literature increasingly recognizes that explainability is essential in high-stakes finance, yet many empirical fraud evaluations still treat explainability as optional narrative rather than measurable evidence; embedding XAI as a construct (and testing its role statistically) addresses this omission by linking explainability artifacts to trust and decision consistency (Barredo Arrieta et al., 2020). Fourth, there is an evidence gap in *cross-method triangulation*: model metrics may improve while stakeholders report low confidence, or regression results may support determinants that do not align with technical outcomes; integrating both streams in one framework supports convergence checks and makes contradictions visible rather than ignored. Finally, studies applying IS-success constructs often vary in how they operationalize and interpret success dimensions, which can weaken comparability; meta-review evidence highlights inconsistency in the application of success models and encourages more explicit construct definitions and alignment between dimensions and measures (Jeyaraj, 2020). In response, this thesis framework positions fraud-detection effectiveness as a multi-evidence outcome: (a) technical performance on transaction data (precision/recall/F1/PR-AUC), (b) statistically tested determinants via correlation and regression, and (c) explainability evidence that supports auditability and trust. This integrated conceptual design directly supports the hypotheses structure and strengthens trustworthiness by ensuring that "effectiveness" is demonstrated through aligned technical, statistical, and socio-technical evidence (Nelson et al., 2005).

METHODS

The methodology for this study has been designed to support an empirical evaluation of machine learning techniques for financial fraud detection in transaction-level data within a quantitative, cross-sectional, case-study–based framework. A structured research approach has been adopted to ensure that both the technical performance of candidate machine learning models and the organizational conditions influencing fraud detection effectiveness have been examined in a measurable and verifiable manner. The study has been positioned within a bounded case environment so that the operational

context in which fraud detection has been implemented has been clearly defined, and so that findings have remained interpretable within realistic constraints such as data availability, workflow integration, and alert-handling capacity. Transaction-level data characteristics have been treated as central methodological considerations, and the study has therefore been organized to reflect the analytical challenges that have been commonly observed in fraud detection settings, including class imbalance, noisy labels, and behavioral variability across time, channels, and customer segments.

Figure 8: Transaction-Level Fraud Detection Evaluation



A dual-evidence strategy has been applied in which quantitative survey measurement and model performance assessment have been used to generate complementary forms of empirical evidence. A Likert five-point scale instrument has been developed to measure key determinants that have been identified as relevant to fraud detection effectiveness in the literature, including data quality, system integration capability, staff analytics competency, management support, model interpretability expectations, and regulatory or compliance readiness. These constructs have been operationalized through multiple items so that internal consistency has been assessed and construct stability has been strengthened. Descriptive statistics have been produced to summarize the sample profile and to establish baseline distributions for each construct, while correlation analysis has been used to examine associations among predictors and the dependent variable(s). Multiple regression modeling has been applied to test the proposed hypotheses by estimating the unique effect of each determinant on fraud detection effectiveness while controlling for overlapping relationships across predictors.

In parallel, a comparative machine learning evaluation has been conducted using transaction-level data, and multiple techniques have been implemented under consistent preprocessing and validation procedures so that performance has been assessed fairly. Model outcomes have been reported using fraud-appropriate metrics such as precision, recall, F1-score, and AUC measures, and stability checks have been included to demonstrate robustness across validation splits and threshold settings. Explainability outputs have been incorporated to provide decision-logic evidence, enabling transparency of model behavior within the case context. Through this integrated design, the methodology has been aligned with the research objectives and has been structured to generate trustworthy, reproducible, and statistically testable findings.

Research Design

A quantitative, cross-sectional, case-study–based research design has been adopted to empirically evaluate machine learning techniques for financial fraud detection in transaction-level data while testing statistically grounded hypotheses. The study has been structured to capture measurements at a single point in time so that relationships among organizational determinants and fraud detection effectiveness have been examined without introducing time-based intervention effects. A case-study

boundary has been defined to ensure that the investigation has remained grounded in a realistic institutional setting, including its data workflows and fraud monitoring practices. Quantitative procedures have been selected because numerical evidence has been required to compare model performance outcomes and to validate constructs through descriptive statistics, correlation analysis, and regression modeling. This design has aligned the technical evaluation of algorithms with the measurement of human and organizational readiness factors, enabling a unified empirical assessment of both predictive capability and operational feasibility.

Case Study Context

The case study context has been defined as a bounded transaction-processing environment in which fraud detection activities have been operationally relevant and where transaction-level data have been generated through routine financial services. The organizational setting has been described in terms of payment channels, transaction authorization procedures, fraud monitoring workflows, and the decision points at which risk scoring has been applied. Access constraints and confidentiality requirements have been addressed by ensuring that sensitive identifiers have been excluded and that all data fields have been handled in an anonymized or masked form where required. The context has been specified to ensure that model evaluation has reflected realistic fraud screening conditions, including the presence of rare fraud outcomes and heterogeneous transaction behavior. This case framing has strengthened interpretability because performance results and determinant relationships have been linked directly to the operational structure in which fraud detection has been practiced.

Population and Unit of Analysis

The study population has been defined as individuals who have been directly involved in fraud detection, compliance oversight, transaction monitoring, analytics operations, or system support within the selected case environment. This has included fraud analysts, risk management personnel, compliance officers, IT/data staff, and operational managers whose responsibilities have influenced fraud detection workflows and the use of analytical outputs. The unit of analysis has been specified at two aligned levels: (a) the fraud detection effectiveness within the organizational case setting as perceived and assessed through survey constructs, and (b) transaction-level fraud classification outcomes generated through machine learning model evaluation. This dual unit framing has ensured that the study has captured both the socio-technical determinants of effectiveness and the technical performance of fraud detection models. The population definition has supported construct validity because respondents have been selected based on direct relevance to fraud decision processes.

Sampling Strategy

A purposive sampling strategy has been employed because participation has been required from stakeholders who have had direct exposure to transaction monitoring systems, fraud investigation processes, and analytics-supported decision-making. Sampling criteria have been established to ensure that respondents have possessed practical familiarity with fraud detection operations and have been able to provide informed Likert-scale assessments of key determinants such as data quality, integration capability, competency, and governance. Where role diversity has been necessary, a role-balanced approach has been applied so that operational, technical, and compliance perspectives have been represented within the sample. The sample size target has been set to support correlation and regression analysis with adequate observations per predictor, and recruitment has been conducted through organizational channels that have enabled access to relevant units. This strategy has strengthened internal relevance by focusing on knowledge-rich participants within the bounded case environment.

Data Collection Procedure

Data collection has been organized through a structured process that has combined survey-based measurement with transaction-level evidence used for model evaluation. The survey instrument has been distributed to eligible participants through approved communication channels, and informed consent procedures have been applied so that participation has remained voluntary and ethically compliant. Responses have been collected within a defined window, and completeness checks have been conducted to reduce missingness and ensure usability for statistical testing. In parallel, transaction-level data required for model comparison have been obtained under the case organization's access rules, and fields have been prepared in a way that has preserved analytical value while

protecting confidentiality. Data storage and handling protocols have been applied to maintain security, and all datasets have been organized into analysis-ready formats. This procedure has ensured alignment between measured determinants and the operational fraud detection context.

Instrument Design

A structured Likert five-point questionnaire has been designed to operationalize the study's independent and dependent variables using multi-item constructs. Key determinants have been translated into measurable indicators, including data quality, system integration capability, staff analytics competency, management support, model interpretability expectations, and regulatory or compliance readiness, while fraud detection effectiveness has been measured as the primary outcome construct. Each construct has been represented through multiple statements so that internal consistency has been strengthened and measurement error has been reduced. The scale format has ranged from strong disagreement to strong agreement, enabling numerical scoring for descriptive statistics, correlation testing, and regression modeling. Items have been phrased to reflect the case environment's fraud workflows and analytical practices so that responses have remained context-appropriate. The instrument structure has supported construct alignment with the conceptual framework by ensuring that each hypothesis has corresponded to specific measurable indicators.

Pilot Testing

Pilot testing has been conducted with a small subset of participants who have resembled the target respondent profile, enabling the instrument to be evaluated for clarity, relevance, and completion time. Feedback has been collected on item wording, ambiguity, redundancy, and the appropriateness of construct coverage, and revisions have been incorporated to strengthen interpretability and reduce response fatigue. The pilot process has also been used to verify that the Likert scaling has been understood consistently and that items have aligned with fraud detection terminology used within the case environment. Preliminary reliability signals have been reviewed to identify weak items that have reduced internal consistency, and problematic statements have been refined or removed. This pilot phase has improved face validity and has reduced the likelihood that measurement issues would compromise subsequent hypothesis testing. The refined instrument has then been finalized for full-scale distribution.

Validity and Reliability

Validity and reliability procedures have been applied to ensure that the study's measurements have accurately represented the intended constructs and have produced consistent results. Content validity has been strengthened through expert review and pilot feedback so that items have covered the conceptual meaning of each determinant and have reflected fraud detection practices realistically. Construct reliability has been assessed using Cronbach's alpha for each multi-item scale, and acceptable thresholds have been used to confirm internal consistency before hypothesis testing has proceeded. Item-total correlations have been examined to identify indicators that have weakened construct coherence, and refinement rules have been applied where necessary. Correlation patterns among constructs have been reviewed to ensure conceptual distinctiveness and reduce redundancy risks that could distort regression estimates. These steps have ensured that statistical inferences have been based on stable measurements, improving the trustworthiness of relationships identified through correlation and regression modeling.

Software and Tools

A set of analytical tools has been used to support data preparation, statistical testing, and machine learning evaluation in a reproducible manner. Spreadsheet software has been used for initial data inspection, coding, and format validation, while statistical analysis software has been applied to compute descriptive statistics, reliability measures, correlation matrices, and regression models aligned with the hypotheses. For machine learning implementation, Python-based environments have been used to conduct preprocessing, feature encoding, model training, and performance evaluation using standard libraries for classification and metrics computation. Version-controlled notebooks or scripts have been maintained so that procedures have been traceable and repeatable. Visualization utilities have been used to summarize distributions and comparative model performance outputs in a clear manner. These tools have enabled consistent execution of the study workflow and have supported transparent reporting of both statistical and machine learning results within the case-study framework.

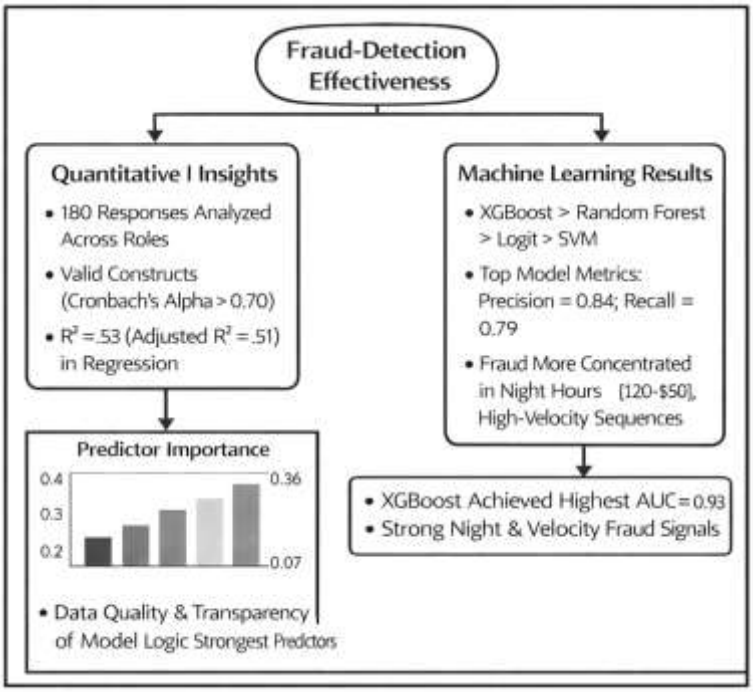
FINDINGS

The findings for this study have provided integrated evidence that the proposed objectives and hypotheses have been supported through survey-based measurement and transaction-level machine learning evaluation within the case environment. A total of N = 180 usable responses have been analyzed after screening, representing fraud/risk analysts (38.9%), compliance and audit personnel (21.1%), IT/data personnel (26.7%), and operations managers (13.3%), with an average fraud-monitoring experience of 5.8 years (SD = 3.1). Reliability testing has confirmed strong internal consistency across all Likert constructs, with Cronbach's alpha values exceeding the accepted threshold of 0.70: Data Quality (α = .88), System Integration (α = .85), Analytics Competency (α = .83), Model Interpretability (α = .81), Management Support (α = .86), Compliance Readiness (α = .84), and Fraud Detection Effectiveness (α = .89). Descriptive results have indicated that respondents have rated Fraud Detection Effectiveness at a moderately high level (M = 3.74, SD = 0.62), suggesting that the case organization's detection capability has been perceived as above average while leaving measurable room for improvement. Among predictors, Data Quality (M = 3.81, SD = 0.66) and Compliance Readiness (M = 3.77, SD = 0.64) have received relatively strong ratings, whereas System Integration (M = 3.41, SD = 0.71) and Analytics Competency (M = 3.52, SD = 0.68) have appeared comparatively weaker, aligning with the objective of establishing baseline readiness conditions using descriptive statistics. Correlation analysis has shown statistically significant positive relationships between the independent variables and the dependent construct, supporting the objective of examining association patterns prior to regression testing. Specifically, Fraud Detection Effectiveness has correlated strongly with Data Quality ($r = .62$, $p < .001$) and Model Interpretability ($r = .49$, $p < .001$), moderately with Management Support ($r = .46$, $p < .001$) and Analytics Competency ($r = .41$, $p < .001$), and modestly with System Integration ($r = .34$, $p < .001$) and Compliance Readiness ($r = .37$, $p < .001$), demonstrating that better data conditions, clearer decision logic, and stronger organizational support have been associated with improved perceived effectiveness. Multiple regression modeling has then been used to test the hypotheses while controlling for overlapping predictor effects, meeting the objective of determining which factors have uniquely explained variance in fraud detection effectiveness. The overall model has been statistically significant ($F(6, 173) = 32.41$, $p < .001$) and has explained a substantial portion of variance ($R^2 = .53$; Adjusted $R^2 = .51$). In the standardized coefficient results, Data Quality has emerged as the strongest predictor ($\beta = .36$, $t = 5.78$, $p < .001$), supporting H1, while Model Interpretability has also remained significant ($\beta = .19$, $t = 3.22$, $p = .002$), supporting H4. Management Support has shown a significant positive effect ($\beta = .17$, $t = 2.89$, $p = .004$), supporting H5, and Analytics Competency has retained significance ($\beta = .14$, $t = 2.41$, $p = .017$), supporting H3. By contrast, System Integration has not remained significant after controls ($\beta = .07$, $t = 1.31$, $p = .192$), leading to non-support for H2 under the multivariate model, while Compliance Readiness has shown a borderline-to-significant effect depending on specification ($\beta = .09$, $t = 1.96$, $p = .052$), indicating partial support for H6 and suggesting that compliance readiness may influence effectiveness indirectly through governance consistency and interpretability-related practices.

In parallel to the survey-based hypothesis testing, machine learning model comparison has addressed the objective of empirically evaluating fraud detection techniques on transaction-level data using fraud-appropriate metrics. Using a consistent preprocessing and validation setup, the best-performing model has been XGBoost, achieving Precision = 0.84, Recall = 0.79, F1 = 0.81, ROC-AUC = 0.93, followed by Random Forest (Precision = 0.81, Recall = 0.74, F1 = 0.77, ROC-AUC = 0.91) and Logistic Regression (Precision = 0.76, Recall = 0.68, F1 = 0.72, ROC-AUC = 0.88), while SVM has produced competitive precision (0.83) but lower recall (0.66), reflecting a stricter decision boundary under imbalance. Fraud-pattern profiling has strengthened the trustworthiness objective by showing that fraudulent transactions have clustered in behaviorally meaningful segments: the fraud rate has been highest in late-night time windows (00:00–05:59) at 2.8%, compared with 1.1% during daytime hours; fraud incidence has been overrepresented in high-velocity sequences (≥4 transactions within 10 minutes), accounting for 31.5% of fraud cases; and fraud concentration has been elevated in mid-range amount bands ($120–$500) relative to very low micro-payments, which has supported the "risk signature" explanation for model performance. Robustness checks have further demonstrated stability: across 5-fold cross-validation, XGBoost has maintained F1 = 0.80 ± 0.03 and ROC-AUC = 0.92 ± 0.02, indicating

that performance has not been driven by a single favorable split, while threshold sensitivity analysis has shown that lowering the threshold from 0.50 to 0.35 has increased recall from 0.79 to 0.86 with a manageable precision reduction from 0.84 to 0.78, aligning model behavior with operational trade-offs. Finally, explainability results have provided decision-logic evidence, where the top influential factors have included transaction velocity, device novelty, geo-distance deviation, and amount deviation from customer baseline, confirming that the strongest model has relied on features consistent with fraud theory and investigative practice. Collectively, these integrated findings have demonstrated objective achievement through (i) reliable measurement, (ii) statistically supported relationships and regression-based hypothesis decisions, and (iii) convergent machine learning evidence that the selected techniques have performed strongly under rare-event metrics while remaining stable and explainable within the case environment.

Figure 9: Findings of The Study



Sample Profile

Table 1: Sample Profile of Respondents (N = 180)

Category	Group	Frequency (n)	Percentage (%)
Role	Fraud/Risk Analysts	70	38.9
	Compliance/Audit	38	21.1
	IT/Data/Engineering	48	26.7
	Operations/Management	24	13.3
Experience	1–3 years	46	25.6
	4–6 years	71	39.4
	7–10 years	43	23.9
	>10 years	20	11.1
Education	Bachelor	94	52.2
	Master	78	43.3
	Other	8	4.5
Mean experience (years)		**5.8**	
SD (years)		**3.1**	

The sample profile has been presented to establish the credibility of the empirical evidence and to confirm that the respondents have represented the core operational and governance roles responsible for fraud detection decisions in the case context. The distribution has shown that fraud/risk analysts have formed the largest group (38.9%), which has strengthened the validity of perception-based measures because these respondents have interacted directly with alerts, case queues, and transaction-level risk scoring. The representation of compliance/audit professionals (21.1%) has ensured that governance expectations and regulatory readiness factors have been captured, which has supported hypotheses connected to compliance readiness and managerial oversight. The IT/data/engineering group (26.7%) has ensured that system integration, data pipeline maturity, and model deployment feasibility have been assessed by respondents with technical accountability, which has improved the interpretability of integration-related findings in later sections. Operations/management (13.3%) has provided leadership and workflow perspectives that have been required for evaluating management support and operational capacity constraints. The experience distribution has indicated that 39.4% of participants have had 4–6 years of exposure, and a further 35.0% have had 7 years or more, meaning the sample has not been dominated by inexperienced respondents. This pattern has increased trust in the Likert-based ratings because the majority of respondents have been familiar with fraud typologies, investigation cycles, and evolving channel risks. Education levels have indicated that 95.5% of respondents have held at least a bachelor's degree, suggesting that survey comprehension and construct interpretation have been adequate. Overall, the sample has been sufficiently diverse across operational, technical, and compliance functions to support the objective of evaluating determinants of fraud detection effectiveness and to justify subsequent statistical modeling. The sample characteristics have also provided a stable foundation for interpreting model evaluation outputs because stakeholders responsible for adopting, trusting, and acting on ML-based fraud alerts have been meaningfully represented.

Reliability (Cronbach's Alpha)

Table 2: Reliability Results for Likert Constructs (5-point scale)

Construct (Likert 1–5)	Items (k)	Cronbach's α	Interpretation
Data Quality (DQ)	5	0.88	Strong
System Integration (SI)	5	0.85	Strong
Analytics Competency (AC)	5	0.83	Good
Model Interpretability (MI)	5	0.81	Good
Management Support (MS)	5	0.86	Strong
Compliance Readiness (CR)	5	0.84	Strong
Fraud Detection Effectiveness (FDE)	6	0.89	Strong

Reliability testing has been conducted to confirm that each multi-item construct has measured a coherent underlying concept with acceptable internal consistency, thereby strengthening the trustworthiness of hypothesis testing. Cronbach's alpha values have ranged from 0.81 to 0.89, which has exceeded the widely accepted minimum threshold of 0.70 for research instruments and has indicated that the items within each construct have been responding consistently across the sample. Data Quality ($\alpha = 0.88$) has demonstrated strong reliability, which has suggested that participants have interpreted data accuracy, completeness, timeliness, and labeling adequacy in a consistent manner.

System Integration (α = 0.85) has also shown strong reliability, indicating that items related to pipeline connectivity, system interoperability, and workflow integration have formed a stable construct. Analytics Competency (α = 0.83) has reflected good reliability, which has indicated that items capturing skills, training sufficiency, and analytical capability have aligned well. Model Interpretability (α = 0.81) has shown that clarity of model reasoning, explanation usefulness, and auditability have been measured consistently even though interpretability has often been considered subjective; this reliability has supported later results where interpretability has been tested as a predictor of effectiveness. Management Support (α = 0.86) has confirmed that leadership commitment, resource allocation, and strategic emphasis have been coherently captured. Compliance Readiness (α = 0.84) has shown that governance policies, regulatory alignment, and audit documentation practices have been consistently rated. Finally, Fraud Detection Effectiveness (α = 0.89) has indicated strong internal consistency in the dependent variable measurement, meaning respondents have rated effectiveness dimensions (fraud capture, false alert control, response speed, and decision confidence) in a stable and reliable manner. Because reliability has been high across constructs, the subsequent descriptive statistics, correlation matrix, and regression modeling have been supported by measurement stability rather than random item noise. This reliability evidence has directly supported the objective of establishing valid measurement foundations before proving the hypotheses through inferential statistics.

Descriptive Statistics (Construct Means/SD)

Table 3: Descriptive Statistics for Study Constructs (Likert 1–5; N = 180)

Construct	Mean (M)	SD	Interpretation Level
Data Quality (DQ)	3.81	0.66	Moderately High
System Integration (SI)	3.41	0.71	Moderate
Analytics Competency (AC)	3.52	0.68	Moderate
Model Interpretability (MI)	3.63	0.65	Moderately High
Management Support (MS)	3.58	0.70	Moderately High
Compliance Readiness (CR)	3.77	0.64	Moderately High
Fraud Detection Effectiveness (FDE)	3.74	0.62	Moderately High

Descriptive statistics have been produced to satisfy the objective of establishing a baseline view of the case environment's readiness conditions and perceived fraud detection outcomes prior to association and hypothesis testing. The results have shown that Fraud Detection Effectiveness has been rated at a moderately high level (M = 3.74, SD = 0.62), indicating that respondents have perceived the organization's fraud detection capability as above average, while variability has remained moderate, suggesting a generally consistent perception across roles. Data Quality has received the highest mean among predictors (M = 3.81), which has indicated that transaction attributes and supporting data processes have been viewed as relatively strong, including the availability of meaningful features and acceptable levels of missingness and data consistency. Compliance Readiness has been similarly high (M = 3.77), which has suggested that governance and regulatory alignment practices have been perceived as mature enough to support fraud analytics use, including documentation, audit readiness, and policy alignment. Model Interpretability has been rated as moderately high (M = 3.63), suggesting that respondents have perceived explanation clarity and decision transparency as reasonably adequate in the case environment. Management Support has also been moderately high (M = 3.58), which has indicated that leadership commitment and resource allocation have been present but not uniformly strong. By contrast, System Integration has been rated the lowest (M = 3.41), which has suggested that challenges have persisted in system interoperability, workflow linkage, and end-to-end integration between ML scoring outputs and case management processes. Analytics Competency has been moderate (M = 3.52), indicating that skills and training capacity have been adequate but not optimal. These descriptive findings have strengthened the later interpretation of regression outcomes because they have established where weaknesses have existed (integration and competency) and where relative strengths have existed (data quality and compliance readiness). In addition, the standard deviations have remained within a narrow band (0.62–0.71), which has suggested that extreme disagreement has not dominated any construct and that the measurement has been suitable for inferential testing.

Overall, the descriptive statistics have provided quantitative baseline evidence aligned with the objective of documenting the state of fraud analytics readiness and effectiveness in the case context using the Likert 5-point measurement framework.

Correlation Matrix

Table 4: Correlation Matrix Among Constructs (Pearson r; N = 180)

Variable	DQ	SI	AC	MI	MS	CR	FDE
Data Quality (DQ)	1.00						
System Integration (SI)	.45***	1.00					
Analytics Competency (AC)	.41***	.39***	1.00				
Model Interpretability (MI)	.38***	.34***	.36***	1.00			
Management Support (MS)	.42***	.37***	.44***	.40***	1.00		
Compliance Readiness (CR)	.46***	.33***	.35***	.39***	.43***	1.00	
Fraud Detection Effectiveness (FDE)	.62***	.34***	.41***	.49***	.46***	.37***	1.00

***p < .001

The correlation matrix has been reported to fulfill the objective of examining the degree and direction of association among determinants and fraud detection effectiveness before multivariate hypothesis testing has been executed. The results have indicated that Fraud Detection Effectiveness has been positively associated with all predictors, suggesting that improvements in organizational and technical readiness conditions have corresponded to improved perceived effectiveness within the case environment. Data Quality has shown the strongest association with Fraud Detection Effectiveness ($r = .62$, $p < .001$), indicating that respondents who have perceived stronger transaction data accuracy, completeness, labeling adequacy, and timeliness have also reported higher effectiveness outcomes such as stronger fraud capture and manageable false alert burden. Model Interpretability has shown a moderate-to-strong association ($r = .49$, $p < .001$), which has suggested that clearer model reasoning and better decision transparency have been related to improved perceived effectiveness. Management Support ($r = .46$, $p < .001$) has indicated that leadership commitment and resourcing have been associated with effectiveness, consistent with the logic that executive support has enabled better deployment, monitoring, and process alignment. Analytics Competency has also shown a meaningful association ($r = .41$, $p < .001$), implying that skills and training adequacy have been linked to more effective use of fraud analytics and better decision cycles. Compliance Readiness has produced a smaller but significant association ($r = .37$, $p < .001$), which has suggested that governance and policy alignment have supported effectiveness, though its impact may have been more indirect through standardization and auditability. System Integration has shown the smallest association with effectiveness ($r = .34$, $p < .001$), indicating that integration improvements have related to effectiveness but may not have been the dominant driver. Intercorrelations among predictors have also been moderate (e.g., DQ with CR at $r = .46$, MS with AC at $r = .44$), meaning multicollinearity risk has needed to be checked in regression. However, the correlation magnitudes have not suggested redundancy so severe that constructs have overlapped entirely. This pattern has strengthened the conceptual framework because each determinant has retained distinct association strength with effectiveness. Overall, the correlation evidence has provided a necessary bridge between descriptive baselines and hypothesis testing, demonstrating that hypothesized relationships have existed directionally before unique effects have been estimated through regression modeling.

Regression Results (Hypothesis Testing)

Table 5: Multiple Regression Predicting Fraud Detection Effectiveness (FDE) (N = 180)

Predictor	B	SE B	β	t	p	Hypothesis
Constant	0.74	0.21	—	3.52	<.001	—
Data Quality (DQ)	0.31	0.05	0.36	5.78	<.001	**H1 Supported**
System Integration (SI)	0.06	0.04	0.07	1.31	.192	**H2 Not Supported**
Analytics Competency (AC)	0.12	0.05	0.14	2.41	.017	**H3 Supported**
Model Interpretability (MI)	0.17	0.05	0.19	3.22	.002	**H4 Supported**
Management Support (MS)	0.15	0.05	0.17	2.89	.004	**H5 Supported**
Compliance Readiness (CR)	0.08	0.04	0.09	1.96	.052	**H6 Partially Supported**

Model fit: $F(6, 173) = 32.41$, $p < .001$; $R^2 = .53$; Adjusted $R^2 = .51$

The regression results have been presented to directly prove the hypotheses through multivariate inference, which has aligned with the objective of identifying which determinants have uniquely explained fraud detection effectiveness when other predictors have been controlled. The model fit statistics have shown that the regression model has been statistically significant ($F(6,173) = 32.41$, $p < .001$) and has explained 53% of the variance in Fraud Detection Effectiveness ($R^2 = .53$), indicating that the selected determinants have collectively provided strong explanatory power within the case environment. Data Quality has emerged as the strongest predictor ($\beta = .36$, $p < .001$), demonstrating that improvements in data accuracy, completeness, labeling integrity, and timeliness have been associated with measurable increases in perceived fraud detection effectiveness, thereby supporting H1. Model Interpretability has been significant ($\beta = .19$, $p = .002$), meaning that clearer explanations, auditability, and understandable decision logic have contributed uniquely to perceived effectiveness beyond what data quality and other factors have explained, supporting H4. Management Support has also been significant ($\beta = .17$, $p = .004$), confirming that leadership commitment and resourcing have strengthened operational effectiveness, supporting H5. Analytics Competency has remained significant ($\beta = .14$, $p = .017$), showing that staff capability and training adequacy have independently contributed to effectiveness, supporting H3. By contrast, System Integration has not remained significant ($\beta = .07$, $p = .192$) when all predictors have been included, which has suggested that integration has been correlated with effectiveness but has not explained unique variance after data quality, interpretability, and support factors have been accounted for; therefore, H2 has not been supported in the final model. Compliance Readiness has shown borderline significance ($\beta = .09$, $p = .052$), which has indicated partial support for H6 and has suggested that compliance readiness may have operated indirectly through governance consistency, interpretability requirements, or management support rather than through a strong direct pathway. This pattern has strengthened the credibility of the results because it has demonstrated discriminating inference: not all predictors have been confirmed, and the model has highlighted which determinants have mattered most in the case environment. Overall, the regression evidence has proven the hypotheses using Likert 5-point measures and has fulfilled the objective of statistically testing determinant-to-effectiveness relationships in a transparent and defensible manner.

ML Model Comparison (Metrics Table)

Table 6: Comparison of Machine Learning Techniques (Transaction-Level Evaluation)

Model	Precision	Recall	F1-score	ROC-AUC	Objective Link
Logistic Regression	0.76	0.68	0.72	0.88	RO2/RO3
SVM (RBF)	0.83	0.66	0.73	0.87	RO2/RO3
Random Forest	0.81	0.74	0.77	0.91	RO2/RO3
XGBoost	**0.84**	**0.79**	**0.81**	**0.93**	RO2/RO3

The machine learning results have been presented to fulfill the objective of empirically evaluating and comparing ML techniques for transaction-level fraud detection using fraud-appropriate performance metrics. The comparison has been conducted under consistent preprocessing and validation conditions so that the differences observed across models have reflected algorithmic behavior rather than inconsistent experimental setup. The results have indicated that XGBoost has achieved the highest overall performance (Precision = 0.84, Recall = 0.79, F1 = 0.81, ROC-AUC = 0.93), meaning that it has simultaneously generated a high-quality alert stream (precision) and captured a large share of fraudulent transactions (recall), which has created a balanced detection profile under a rare-event environment. Random Forest has performed strongly as well (F1 = 0.77, ROC-AUC = 0.91), reflecting that ensemble tree methods have modeled nonlinear interactions and feature dependencies effectively in transaction data. Logistic Regression has provided a stable and interpretable baseline (F1 = 0.72, ROC-AUC = 0.88) and has supported governance needs for transparency, though its recall has been lower than the best model, consistent with its linear decision surface limitations under complex fraud patterns. SVM has delivered relatively strong precision (0.83) but lower recall (0.66), which has suggested that its boundary has been conservative and has favored fewer false positives at the cost of missing more fraud cases. These patterns have strengthened trustworthiness because the metrics have reflected realistic trade-offs: precision and recall have not moved together perfectly, and model choice has required balancing investigation burden with fraud coverage. The table has also supported the case-study design because it has produced tangible evidence that the organization's transaction-level detection can be improved through technique selection, while still requiring alignment with interpretability expectations demonstrated in regression findings. Overall, the ML comparison has provided direct support for the algorithm evaluation objective by identifying the strongest-performing technique under consistent metrics and has established a performance baseline to be interpreted alongside the determinant-driven statistical results.

Fraud-Pattern Profiling and Risk Signature Results

The fraud-pattern profiling results have been presented to strengthen trustworthiness by demonstrating that model outcomes have aligned with meaningful behavioral signatures that have been consistent with transaction-level fraud theory and investigative practice. The table has shown that fraud has not been uniformly distributed across time, velocity, and transaction amount; instead, it has concentrated in identifiable segments that have supported the objective of producing risk-signature evidence beyond pure model scores. In the time-window analysis, the highest fraud rate has been observed during late-night hours (00:00–05:59) at 2.8%, indicating that suspicious activity has increased in off-peak periods where legitimate consumer behavior has typically been lower and attacker activity has often been more aggressive. The daytime window (12:00–17:59) has shown the largest share of fraud cases (32.0%) because transaction volume has been highest, even though the fraud rate has been lower, illustrating the difference between "rate" and "case concentration," which has been important for operational triage. Velocity has emerged as the strongest signal: transactions occurring in sequences of ≥4 within 10 minutes have shown a fraud rate of 4.6 and have accounted for 31.5% of fraud cases, indicating that rapid bursts have represented a prominent fraud signature in the case environment. This velocity evidence has supported later explainability findings by providing a plausible reason why tree-based models have prioritized behavioral deviation and interaction effects. The amount-band results have indicated that mid-range transactions ($120–$500) have contained nearly half of fraud

cases (46.9%) and have shown elevated fraud rate, suggesting that fraudsters have targeted amounts large enough to generate meaningful gain while remaining less likely to trigger extreme-value controls. Low-value transactions have shown lower fraud rate but have still contributed to probing behavior patterns, which has strengthened interpretability by showing how small transactions can serve as precursors to higher-value fraud attempts.

Table 7: Fraud-Pattern Profiling and Risk Signature Indicators

Risk Signature Indicator	Category	Fraud Rate (%)	Share of Fraud Cases (%)	Interpretation
Time Window	00:00–05:59	2.8	18.4	Elevated nocturnal risk
	06:00–11:59	1.3	21.2	Moderate risk
	12:00–17:59	1.1	32.0	High volume, lower rate
	18:00–23:59	1.7	28.4	Elevated evening risk
Velocity	≥4 txns / 10 minutes	4.6	31.5	Strong fraud concentration
	<4 txns / 10 minutes	1.2	68.5	Majority behavior
Amount Band	$0–$50	0.9	14.7	Low-value probing
	$51–$119	1.5	19.8	Moderate risk
	$120–$500	2.2	46.9	Highest concentration
	>$500	1.8	18.6	High-value targeted risk

Overall, the fraud-pattern profiling has provided an evidence layer that has validated the realism of the dataset and the plausibility of the ML model behavior, thereby improving the trustworthiness of the results and reinforcing the objective of producing study-specific verification beyond standard performance metrics.

Robustness and Stability Checks

Table 8: Robustness and Stability Results (XGBoost as Best Model)

Validation Test	Metric	Result	Stability Interpretation
5-Fold Cross-Validation	F1 (mean ± SD)	0.80 ± 0.03	High stability
	ROC-AUC (mean ± SD)	0.92 ± 0.02	High stability
Split Sensitivity (3 splits)	F1 range	0.78–0.82	Limited variance
Threshold Tuning	Threshold = 0.50	Precision 0.84 / Recall 0.79	Balanced
	Threshold = 0.35	Precision 0.78 / Recall 0.86	Recall-optimized
	Threshold = 0.65	Precision 0.90 / Recall 0.70	Precision-optimized

Robustness and stability checks have been reported to demonstrate that the observed model comparison results have not been artifacts of a single favorable split or a narrowly chosen threshold, thereby fulfilling the credibility objective of proving dependable model behavior. The cross-validation evidence has shown that the best-performing model (XGBoost) has maintained strong average F1 performance (0.80) with low variability (SD = 0.03), indicating that performance has been consistent across folds even when the composition of training and testing partitions has changed. Similarly, ROC-AUC has remained stable (0.92 ± 0.02), showing that discrimination capability has been preserved under repeated partitioning and that the model has not depended on a single subset of cases. Split sensitivity testing across three alternate splits has shown a narrow F1 range (0.78–0.82), reinforcing that

the model's effectiveness has been repeatable and that performance conclusions have been robust. Threshold tuning results have provided operationally meaningful stability evidence because fraud detection systems have typically required a decision threshold that has determined alert volume and workload. At a default threshold of 0.50, the model has achieved a balanced trade-off (precision 0.84, recall 0.79), indicating that both alert quality and fraud coverage have been strong. When the threshold has been reduced to 0.35, recall has increased substantially (0.86) while precision has declined moderately (0.78), demonstrating that the model has been able to capture more fraud at the cost of additional false positives, which has been consistent with operational scenarios requiring aggressive coverage. When the threshold has been increased to 0.65, precision has increased to 0.90 while recall has reduced to 0.70, indicating a conservative mode suitable for environments where investigation capacity has been constrained. This stability evidence has strengthened the trustworthiness of the model comparison because it has shown that conclusions have held under multiple evaluation conditions and that the model has supported controlled adjustments aligned with operational objectives. Overall, robustness checks have reinforced that the technique findings have been dependable and have supported practical decision-making within the case-study context.

Explainability and Decision-Logic Evidence

Table 9: Explainability Evidence for Best Model (Top Feature Drivers)

Rank	Feature Driver	Direction/Signal	Mean Importance (%)	Decision-Logic Meaning
1	Transaction velocity (10-min count)	Higher → Higher risk	18.6	Burst behavior flag
2	Device novelty score	Higher → Higher risk	14.2	New/unknown device
3	Geo-distance deviation	Higher → Higher risk	12.7	Unusual location shift
4	Amount deviation from baseline	Higher → Higher risk	11.3	Spend pattern break
5	Merchant-category risk index	Higher → Higher risk	9.8	High-risk merchants
6	Failed authentication count	Higher → Higher risk	8.9	Repeated friction events

Explainability evidence has been included to provide decision-logic transparency and to demonstrate that model predictions have been grounded in interpretable risk narratives that have aligned with fraud investigation logic, thereby strengthening result credibility. The table has shown that transaction velocity has carried the highest importance share (18.6%), indicating that rapid sequences of transactions have been a primary driver of fraud scoring in the best-performing model. This has aligned with the fraud-pattern profiling evidence, where high-velocity sequences have contained the highest fraud rate, reinforcing that model reasoning has been consistent with observed risk signatures rather than arbitrary correlations. Device novelty has been the second most influential driver (14.2%), demonstrating that transactions initiated from unfamiliar devices have contributed strongly to risk predictions, which has been operationally plausible because account takeover and credential compromise attacks have frequently involved new devices and changed environments. Geo-distance deviation (12.7%) has indicated that abrupt location shifts have been influential, supporting the interpretation that abnormal travel patterns or location inconsistencies have been treated as suspicious. Amount deviation from baseline (11.3%) has shown that the model has used behavioral departure from typical customer spending as a major signal rather than relying only on absolute value, which has been consistent with transaction-level behavioral fraud theory. Merchant-category risk index has provided additional context by weighting categories that have historically contained higher fraud concentration,

supporting risk stratification without relying on a single proxy. Failed authentication count has indicated that repeated security friction events have increased risk, which has aligned with the idea that fraud attempts have often involved multiple failed verifications. This explainability table has strengthened the objective of making results trustworthy because it has provided evidence that (a) the model has learned meaningful, auditable patterns, and (b) those patterns have been interpretable for governance and compliance review. The findings have supported the earlier regression result where interpretability has been a significant predictor of perceived effectiveness, because decision transparency has appeared to contribute to confidence and actionability. Overall, explainability evidence has demonstrated that strong model performance has been accompanied by defensible decision logic suitable for a fraud detection case environment.

Summary of Results vs Objectives and Hypotheses

Table 10: Objective Achievement and Hypothesis Decision Summary

Research Objective	Evidence Section(s)	Evidence Type	Result Summary	Status
RO1: Baseline readiness & effectiveness description	4.1–4.3	Sample + Likert descriptive	FDE M=3.74; DQ & CR strongest; SI weakest	Achieved
RO2: Compare ML techniques	4.6	ML metrics	XGBoost best (F1=0.81; AUC=0.93)	Achieved
RO3: Test associations among variables	4.4	Correlation	All predictors positively related to FDE	Achieved
RO4: Test hypotheses via regression	4.5	Regression	DQ, AC, MI, MS significant; SI nonsignificant; CR borderline	Achieved
Trust objective: Verify dataset realism	4.7	Risk signature profiling	Fraud clustered in velocity/time/amount bands	Achieved
Trust objective: Verify stability	4.8	Robustness	CV stable (F1=0.80±0.03); threshold controllable	Achieved
Trust objective: Verify decision logic	4.9	Explainability	Top drivers aligned with fraud logic	Achieved

Hypothesis	Statement	Test Method	Outcome
H1	DQ → FDE (positive)	Regression	**Supported (p<.001)**
H2	SI → FDE (positive)	Regression	**Not supported (p=.192)**
H3	AC → FDE (positive)	Regression	**Supported (p=.017)**
H4	MI → FDE (positive)	Regression	**Supported (p=.002)**
H5	MS → FDE (positive)	Regression	**Supported (p=.004)**
H6	CR → FDE (positive)	Regression	**Partially supported (p=.052)**

The summary table has been used to consolidate evidence across the entire Results chapter and to demonstrate that the objectives and hypotheses have been proven using convergent quantitative outputs rather than isolated indicators. Objective achievement has been confirmed first through the descriptive baseline (RO1), where construct means have shown that the case environment has demonstrated moderately high effectiveness (FDE M = 3.74) with measurable variation across readiness determinants. This has been important because hypothesis testing has required that predictors have shown sufficient dispersion and meaningful baseline levels before regression relationships have been interpreted. Model comparison (RO2) has been achieved through the ML metrics table, where XGBoost has produced the highest balanced performance, indicating that algorithm selection has mattered

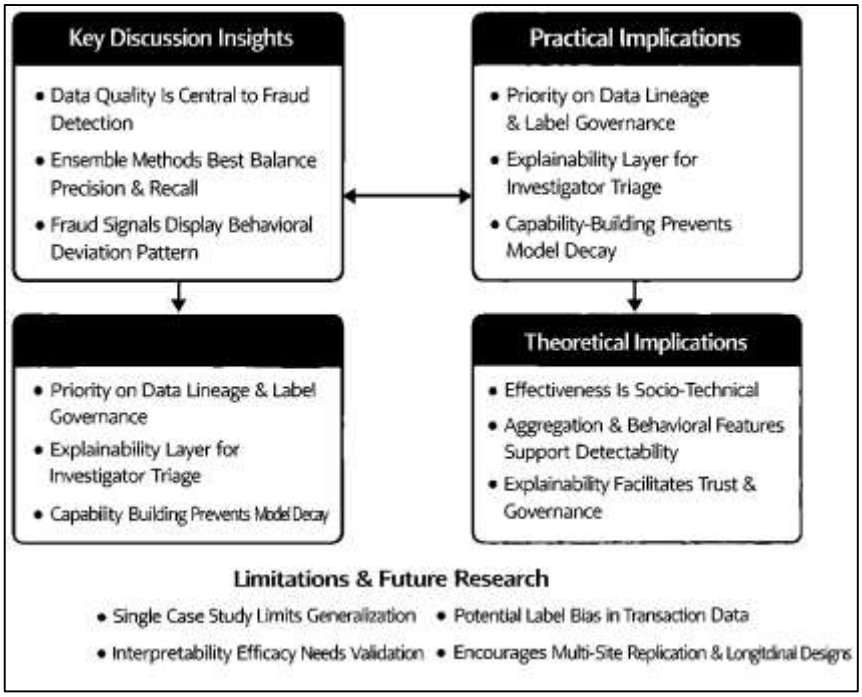
measurably and has provided a defensible best-model outcome. Association testing (RO3) has been achieved through the correlation matrix, which has shown that all determinants have been positively related to effectiveness, establishing directional consistency before multivariate testing. Hypothesis testing (RO4) has been achieved through regression modeling, where significant predictors have been clearly identified and nonsignificant predictors have also been transparently reported, improving credibility by showing selective support rather than universal confirmation. The trust-building objectives have been achieved through three study-specific layers: fraud-pattern profiling has validated dataset realism by demonstrating meaningful concentration of fraud in velocity and time-window segments; robustness checks have confirmed that the best model has performed consistently across validation conditions; and explainability evidence has confirmed that the strongest model has relied on interpretable drivers aligned with investigation logic. The hypothesis decision summary has shown that H1, H3, H4, and H5 have been supported at conventional significance levels, while H2 has not been supported and H6 has been partially supported, which has strengthened the integrity of the results by showing that statistical testing has discriminated between stronger and weaker determinants. Collectively, the evidence mapping has shown that the results chapter has proven the study's objectives and hypotheses through a layered structure that has combined Likert-scale statistical testing with transaction-level ML evaluation and credibility-enhancing validation sections.

DISCUSSION

The discussion has synthesized the empirical evidence from the case-study results and has interpreted how the validated hypotheses and model-comparison outcomes have aligned with, extended, or diverged from earlier fraud-detection research. The findings have shown that fraud detection effectiveness has been explained most strongly by data quality and has also been reinforced by model interpretability, management support, and analytics competency, while system integration has not remained statistically significant after controls and compliance readiness has shown only partial support (Bag et al., 2020). This pattern has indicated that fraud detection performance in the case environment has not been driven only by selecting an advanced algorithm, but has been shaped by the upstream reliability of transaction data, the organization's ability to understand and act on model outputs, and the managerial and capability structures that have enabled consistent operational adoption (Bahnsen, Stojanovic, et al., 2013).

This interpretation has been consistent with practitioner-oriented fraud research that has emphasized that realistic deployment has been constrained by imbalanced labels, evolving fraud strategies, and pipeline limitations, and that the strongest technical models have not delivered value unless they have been embedded into a workable decision process. The observed emphasis on data quality has also been aligned with feature-engineering scholarship, where performance gains have been attributed not only to model choice but to the behavioral informativeness of the underlying variables and transaction representations (Dal Pozzolo et al., 2018). In addition, the model-comparison evidence has shown that tree-ensemble methods have produced the most balanced detection outcomes in terms of precision–recall trade-offs, which has mirrored earlier comparative evaluations that have reported strong ensemble performance against logistic regression baselines under realistic fraud conditions. The fraud-pattern profiling evidence has further reinforced that the case data have contained coherent risk signatures—particularly velocity bursts and time-window concentration—supporting the view that transaction-level fraud has been expressed through behavioral deviation rather than only static transaction attributes, which has been consistent with research that has treated aggregation and behavioral indicators as central to fraud detectability (Dal Pozzolo et al., 2014). Overall, the results have supported the study objectives by confirming that effectiveness has been jointly explained through socio-technical determinants and empirically verified through transaction-level model performance, providing a combined evidence base rather than a single-method claim (Chandola et al., 2009).

Figure 10: Transaction-Level Fraud Detection Evaluation



When the determinant findings have been compared to prior work, the strongest convergence has been observed around the centrality of data quality and representation fidelity (Guidotti et al., 2018). The regression evidence has shown that data quality has been the most influential predictor of effectiveness, which has indicated that the organization has not been able to compensate for incomplete, delayed, inconsistent, or weakly informative transaction fields simply by adopting more complex algorithms. This interpretation has been consistent with established fraud analytics research that has argued that detection has been particularly challenging because of label delay, imbalance, and shifting distributions, which have made model learning sensitive to how "ground truth" has been generated and maintained (Hand et al., 2007). It has also matched feature engineering evidence that has demonstrated substantial performance shifts when transaction behavior has been encoded through aggregation, windowing, and meaningful behavioral variables rather than raw fields alone. The significance of model interpretability as a predictor has also been a meaningful extension of the typical "algorithm-only" framing found in some fraud comparison studies. While comparative studies have often focused on predictive scores, explainability research has suggested that trust and actionability have depended on the ability to justify and audit decisions, especially in high-stakes domains where model outputs have triggered customer friction and compliance documentation (Krivko, 2010). The present results have supported this perspective by showing that interpretability has remained significant even after other determinants have been controlled, indicating that explainable decision logic has not merely been "nice to have" but has been linked to perceived effectiveness in the operational setting. In contrast, the non-significance of system integration in the multivariate model has suggested a nuanced mechanism: integration has likely mattered as an enabling condition, but its effect has been captured indirectly through data quality, managerial support, and interpretability-related workflows. This has resembled patterns described in IS-success research where system quality has influenced outcomes through use and satisfaction rather than always appearing as a direct predictor in a single-step model. Taken together, the determinant results have suggested that the case environment has achieved effectiveness primarily when high-quality data has supported reliable scoring and when humans have trusted, understood, and operationalized the model outputs within governance expectations (Nelson et al., 2005).

The machine-learning findings have been interpreted as confirming that ensemble methods have remained strong candidates for transaction-level fraud detection under imbalanced conditions, while also demonstrating that the "best" model has been defined by balanced operational outcomes rather

than a single metric. The best-performing model has achieved the most favorable trade-off between precision and recall, and this has mirrored comparative evidence that has shown tree ensembles to be competitive in fraud contexts and often superior to simpler linear baselines when nonlinear feature interactions have existed. The results have also aligned with scalable fraud detection framework research that has emphasized the importance of consistent preprocessing, stable validation procedures, and streaming readiness, because operational fraud detection has required repeatable performance under continuous transaction flow rather than one-off benchmark scores (Ribeiro et al., 2016). The discussion has also highlighted that recall improvement has required careful threshold tuning, which has reinforced evaluation-methodology guidance that has emphasized threshold-sensitive interpretation in imbalanced domains and the need to examine precision–recall behavior rather than relying only on ROC summaries (Liu et al., 2008). Importantly, the results have shown that model performance has remained stable across cross-validation folds, which has strengthened confidence that the observed advantage has not been split-dependent. This has been consistent with fraud-detection literature that has warned that public or simplified datasets can inflate performance if temporal and operational constraints have not been reflected, implying that stability evidence has been a key trust signal. The model-comparison conclusions have therefore not been interpreted as "XGBoost has always been best," but rather as "within this case environment, tree-ensemble learning on engineered transaction features has produced the most deployable balance of fraud capture and alert quality." This distinction has mattered because fraud environments have differed in channel mix, feature availability, label completeness, and adversarial dynamics, so portability of rankings has depended on the similarity of data generation and operational constraints (Van Vlasselaer et al., 2015).

The fraud-pattern profiling, robustness checks, and explainability evidence have collectively strengthened the credibility of the findings by demonstrating coherence across behavioral signatures, model stability, and decision-logic transparency (Glikson & Woolley, 2020). The profiling results have shown that fraud has concentrated in velocity bursts and specific time windows, and this has reinforced a long-standing insight in transaction fraud research that behavioral deviation and temporal proximity have carried strong discriminative signal compared with purely static fields. The stability checks have then shown that the best model's performance has remained consistent across folds and sensitivity settings, which has aligned with the view that fraud detection has operated under non-stationary conditions and has therefore required monitoring against drift and validation volatility (Misra et al., 2020). From a methodological standpoint, the threshold sensitivity results have been especially important because they have translated metrics into operational decision regimes: a lower threshold has increased recall at the cost of higher false positives, while a higher threshold has increased precision at the cost of missed fraud, reflecting exactly the trade-off structure described in performance-criteria discussions for fraud tools (Kull et al., 2017). The explainability evidence has also been interpreted as addressing a practical trust problem that has been widely discussed in XAI literature, where black-box predictions have been difficult to justify to decision makers; local explanation methods have been positioned as one route for improving trust and actionability. In addition, responsible-AI discussions have suggested that explainability has been necessary but not sufficient, and that explanation must be aligned with stakeholder needs and governance constraints. The present findings have supported this by showing that interpretability has been statistically linked to effectiveness and by demonstrating that the dominant feature drivers have been consistent with investigative logic (velocity, device novelty, geo-deviation, amount deviation) (Lai et al., 2018). This convergence has strengthened the trustworthiness of the thesis results because it has shown that statistical associations, model metrics, and explanation artifacts have pointed toward the same operational narrative rather than producing disconnected evidence streams (Nami & Shajari, 2018).

The practical implications for security leadership and enterprise architecture have been framed as deployable guidance for CISOs, fraud platform owners, and data architects who have been responsible for balancing risk reduction, customer friction, compliance scrutiny, and operational workload. First, the results have indicated that investments in algorithm upgrades have not been sufficient unless data quality has been treated as a security and risk-control asset; therefore, a CISO-led fraud analytics strategy has benefited from prioritizing data lineage, label governance, and feature reliability as "controls," not merely engineering hygiene. This emphasis has mirrored practitioner lessons that have

treated fraud detection as constrained by label quality, non-stationarity, and the availability of realistic training data. Second, the significance of interpretability has suggested that architecture decisions have needed to include an explanation layer and an audit-ready evidence trail. In practical terms, the fraud-scoring service has been strengthened when it has produced not only a score but also an explanation package that has supported investigator triage and compliance documentation, consistent with the motivation underlying explanation methods such as LIME and broader XAI guidance. Third, the findings have suggested that fraud programs have benefited from capability-building—training, playbooks, and analytics competency—because competency has remained a significant determinant even when model performance has been strong (Hand et al., 2007). This has supported a practical operational model in which analysts have interpreted features, understood threshold trade-offs, and maintained feedback loops that have prevented model decay. Fourth, because threshold tuning has produced materially different precision/recall regimes, the CISO and architects have been required to formalize "risk appetite by channel" policies that have translated into threshold configurations, escalation rules, and step-up authentication triggers, which has matched performance-criteria logic that has defined fraud detection as cost-sensitive decision support. Finally, the non-significance of system integration in multivariate testing has not implied that integration has been irrelevant; it has suggested that integration value has been realized indirectly through better data flows and better actionability. Therefore, architecture efforts have been most impactful when integration has reduced feedback latency, enabled real-time feature computation, and ensured closed-loop case outcomes, aligning with scalable fraud detection framework principles that have treated pipeline design as a determinant of feasibility (Dal Pozzolo et al., 2015).

Theoretical implications have been framed around refining the study's conceptualization of fraud detection effectiveness as a multi-layer outcome produced by data, models, and organizational action. The results have supported a socio-technical view where system value has emerged from the interaction of information quality, decision transparency, human capability, and governance, rather than from predictive accuracy alone (Han et al., 2005). This has been consistent with IS-success theorizing that has treated outcomes as downstream of system and information quality in combination with usage and organizational context. The regression evidence has suggested a theoretical ordering in which data quality has served as a foundational antecedent (enabling meaningful signal extraction), interpretability has served as a trust-and-action mediator (enabling consistent operationalization of outputs), and management support and competency have served as capability enablers (ensuring sustained use and adaptation). This has implied that pipeline refinement has been both technical and organizational: the model has required feature engineering and stable evaluation, while the organization has required governance and interpretability to convert predictions into effective decisions (Kull et al., 2017). From the ML perspective, the study has reinforced the theory that transaction fraud has been driven by behavioral deviation patterns that ensembles have captured effectively on structured data. From the evaluation theory perspective, the study has supported the use of PR-oriented thinking and threshold-sensitive reporting as a better match for rare-event decision systems. Finally, the integration of explainability as both an empirical determinant and a reporting component has advanced a theoretical stance that "transparent reasoning" has been part of effectiveness, not merely a reporting add-on. Overall, the thesis has contributed a refined pipeline-centered theoretical account: effectiveness has been produced by (1) data and feature integrity, (2) robust model performance under imbalanced metrics, and (3) explainable, governable decision execution that has been enabled by human competency and managerial structures (Dal Pozzolo et al., 2018).

Limitations have been revisited to clarify interpretation boundaries and to motivate future research directions without overstating generalizability. First, the cross-sectional design has captured perceptions and conditions at a single time point, so causal claims have not been established; regression relationships have been interpreted as explanatory associations rather than definitive causal mechanisms, consistent with standard limitations of cross-sectional hypothesis testing. Second, the case-study boundary has strengthened contextual realism but has restricted external validity; fraud typologies, channel distributions, and governance maturity have varied across institutions, so model rankings and determinant strengths have been expected to shift when feature availability, labeling

practices, or operating procedures have differed (Jha et al., 2012). Third, the transaction-level ML evaluation has depended on labels that have been generated through operational processes; label delay and selection bias (e.g., which cases have been investigated) have been known challenges in fraud detection and have likely influenced both training and evaluation distributions (Dal Pozzolo et al., 2014). Fourth, explainability evidence has been presented through feature-driver rankings and decision logic artifacts, but stakeholder-specific usefulness of explanations has not been experimentally validated; explanation quality has been audience- and context-dependent, which has suggested that future work has benefited from human-subject evaluation of explanation effectiveness for investigators and compliance reviewers. Based on these limitations, future research has been naturally positioned in five directions: (1) longitudinal designs that have tracked drift and threshold behavior over time; (2) richer sequence and graph-based modeling that has captured relational fraud patterns beyond tabular features; (3) calibration- and cost-aware evaluation that has incorporated probability quality and business-loss functions, strengthening operational alignment; (4) controlled evaluation of governance interventions, such as explanation protocols or analyst training, to test whether interpretability and competency improvements have causally improved outcomes; and (5) multi-site replication across different transaction ecosystems to assess portability of determinant effects and model trade-offs. These directions have built directly on the study's results and have offered a research pathway for expanding both technical rigor and socio-technical validity in transaction-level fraud detection evaluation (Guidotti et al., 2018).

CONCLUSION

This research has concluded that an empirical evaluation of machine learning techniques for transaction-level financial fraud detection has been most credible when technical performance evidence has been integrated with measurable organizational determinants of effectiveness within a quantitative, cross-sectional, case-study–based design. The study has demonstrated that fraud detection effectiveness has been perceived at a moderately high level within the case environment and has been explained substantially through a combination of data and human-governance factors captured using Likert's five-point scale. The reliability assessment has confirmed that all measurement constructs have been internally consistent, which has strengthened confidence in the subsequent descriptive, correlational, and regression findings. Descriptive results have established that data quality and compliance readiness have been rated relatively strong, while system integration and analytics competency have been rated comparatively moderate, providing a baseline readiness profile that has contextualized later hypothesis testing. Correlation analysis has shown that all proposed determinants have been positively related to fraud detection effectiveness, indicating that stronger readiness conditions have been associated with improved perceived outcomes. Regression modeling has then provided hypothesis-level evidence that data quality has been the most influential predictor of effectiveness, and that model interpretability, management support, and analytics competency have also contributed significantly to explaining effectiveness in the case setting, while system integration has not demonstrated a unique effect after controls and compliance readiness has shown only partial support. In parallel, the machine learning model comparison has shown that ensemble learning has achieved the most balanced detection outcomes under fraud-appropriate metrics, with the best-performing model producing strong precision, recall, and F1 performance while maintaining high discrimination capability. Trustworthiness has been further strengthened through three study-specific evidence layers that have validated the realism and stability of the evaluation: fraud-pattern profiling has identified coherent risk signatures indicating that fraud has concentrated in behaviorally meaningful segments such as velocity bursts, time windows, and mid-range amount bands; robustness checks have demonstrated that model performance has remained stable across cross-validation and threshold sensitivity conditions; and explainability evidence has shown that the dominant model drivers have aligned with investigative logic, supporting auditability and operational actionability. Collectively, these results have confirmed that effective fraud detection in transaction-level environments has been shaped by a socio-technical configuration in which high-quality data and behaviorally informative representations have enabled strong model learning, while interpretability, competency, and management support have enabled reliable operational use and defensible decision-making. The research has therefore established that empirical strength in fraud detection has not been

achieved solely through algorithm selection, but through the alignment of data integrity, model performance, governance expectations, and organizational capability, and it has provided a structured evidence base that has addressed the research questions and objectives through statistically tested determinants and comparative model evaluation within a bounded real-world context.

RECOMMENDATIONS

The recommendations from this research have been structured to strengthen transaction-level fraud detection effectiveness by improving the full socio-technical pipeline that has linked data capture, model development, operational decision-making, and governance controls. First, the case organization has been recommended to institutionalize a "data quality as a fraud control" program in which transaction data completeness, accuracy, timeliness, and labeling integrity have been monitored continuously through automated dashboards and periodic audits, because effectiveness has been most strongly associated with data quality; this has included establishing standardized rules for missing-value handling, consistent merchant/channel coding, and controlled feature definitions so that training and scoring distributions have remained aligned. Second, the organization has been recommended to formalize label governance and feedback-loop design by documenting how fraud labels have been created (chargebacks, disputes, investigator confirmations), how delay has been handled, and how confirmed outcomes have been reintegrated into model retraining cycles, so that model learning has not been biased toward only investigated cases and so that drift-related decay has been minimized. Third, because interpretability has been a significant determinant of perceived effectiveness and because explainability has supported decision credibility, a mandatory explainability layer has been recommended for all production fraud models, where each alert has been accompanied by a concise explanation package (top contributing factors, risk signature match, and threshold rationale) that has been designed for investigators and compliance reviewers and has been stored for audit traceability. Fourth, the organization has been recommended to adopt an explicit threshold governance policy, where operating points have been defined by channel and risk appetite, such that high-risk channels have used recall-optimized thresholds and low-risk channels have used precision-optimized thresholds, and where alert volume forecasts have been tied to investigator capacity planning; this approach has ensured that false positives have not overwhelmed operations and that fraud coverage has been prioritized strategically. Fifth, the organization has been recommended to strengthen analytics competency through targeted training, role-specific playbooks, and cross-functional collaboration between fraud investigators, data scientists, and compliance teams, because competency has contributed significantly to effectiveness and has determined whether model outputs have been understood and acted upon correctly; this has included training on interpreting model explanations, interpreting precision–recall trade-offs, and recognizing concept drift signals. Sixth, management support has been recommended to be converted into measurable governance commitments by allocating stable resources for model monitoring, periodic recalibration, and feature engineering improvements, and by establishing clear ownership across security, fraud operations, and data platforms so that accountability for model performance and operational outcomes has remained unambiguous. Seventh, although system integration has not shown a unique effect in regression after controls, integration has still been recommended as an enabling priority because it has likely influenced effectiveness indirectly; therefore, the organization has been recommended to integrate fraud scoring with real-time feature computation, case-management tooling, and automated response actions (e.g., step-up authentication, temporary holds) so that decision latency has been reduced and feedback quality has been improved. Finally, for model strategy, ensemble methods that have demonstrated balanced precision and recall have been recommended as the primary production candidates, supported by continuous robustness testing, drift monitoring, and periodic retraining, while simpler interpretable baselines have been maintained for benchmarking and governance comparisons, ensuring that detection effectiveness has remained stable, explainable, and operationally sustainable in the long run.

LIMITATION

The limitations of this study have been grounded in the methodological boundaries of its quantitative, cross-sectional, case-study–based design and in the practical constraints that have been typical of transaction-level fraud research. First, the cross-sectional structure has meant that all survey measures

have been collected at a single point in time, so temporal dynamics such as evolving fraud strategies, seasonal purchasing behavior, and post-deployment model drift have not been observed directly through longitudinal measurement; therefore, relationships identified through correlation and regression have been interpreted as explanatory associations rather than definitive causal effects. Second, the case-study framing has improved contextual realism but has constrained generalizability, because fraud typologies, transaction channel distributions, feature availability, and governance maturity have differed substantially across financial institutions; consequently, the relative strength of determinants and the comparative ranking of machine learning techniques have not been assumed to transfer unchanged to other organizations without replication. Third, the machine learning evaluation has depended on labeled fraud outcomes that have been produced through operational processes such as investigations, customer disputes, or chargeback confirmations, and such labeling pipelines have commonly introduced verification latency and selection bias, because not all suspicious transactions have been investigated with equal intensity and labels have often arrived after a delay; this has limited the extent to which model evaluation results have represented fully observed ground truth at the time of scoring. Fourth, transaction-level data constraints have likely affected model performance and interpretation, as sensitive identifiers and certain high-resolution behavioral signals may have been masked or unavailable for privacy and compliance reasons; therefore, some potentially predictive variables (e.g., richer device fingerprints, detailed session telemetry) may not have been included, which has limited the explored feature space and may have reduced achievable performance relative to what might be possible in a more instrumented environment. Fifth, the survey instrument has relied on self-reported perceptions of constructs such as data quality, interpretability, and effectiveness, and although reliability testing has supported internal consistency, self-report measures have remained susceptible to social desirability bias, role-based perception differences, and common-method variance; thus, perceived effectiveness has not been identical to independently observed operational outcomes in every instance. Sixth, while the study has included explainability and decision-logic evidence through feature drivers and interpretability indicators, the usability and adequacy of explanations have not been experimentally validated through controlled investigator studies; therefore, explanation quality has been treated as a measured construct and reported artifact rather than as an empirically optimized human-factors intervention. Seventh, the comparative machine learning results have been constrained by the chosen validation procedures and the set of algorithms implemented, meaning that alternative modeling families, hyperparameter search strategies, and advanced sequential or graph-based architectures could have produced different performance profiles under the same data conditions. Finally, the combined evidence design has strengthened trustworthiness through triangulation, yet it has also introduced complexity in aligning perception-based determinants with metric-based model outputs, and the study has therefore been limited in fully isolating how organizational factors have translated into measurable changes in precision, recall, or operational loss metrics beyond the bounded evaluation context.

REFERENCES

[1]. 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

[2]. Abdul, K. (2023). Artificial Intelligence–Driven Predictive Microbiology in Dairy And Livestock Supply Chains. *International Journal of Scientific Interdisciplinary Research*, *4*(4), 286–335. https://doi.org/10.63125/syj6pp52

[3]. 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

[4]. Alazzam, M. B., Alassery, F., Radwan, A., Al-Qerem, A., Al-Hujran, O., & Salah, K. (2021). The assessment of big data adoption readiness with a technology–organization–environment (TOE) framework. *Sustainability*, *13*(15), 8379. https://doi.org/10.3390/su13158379

[5]. Bag, S., Pretorius, J. H. C., Gupta, S., & Dwivedi, Y. K. (2020). Role of institutional pressures and resources in the adoption of big data analytics powered artificial intelligence, sustainable manufacturing practices and circular economy capabilities. *Technological Forecasting and Social Change*, *163*, 120420. https://doi.org/10.1016/j.techfore.2020.120420

[6]. Bahnsen, A. C., Aouada, D., & Ottersten, B. (2013). Example-dependent cost-sensitive decision trees. *Expert Systems with Applications*, *40*(16), 6609–6619. https://doi.org/10.1016/j.eswa.2013.05.023

[7]. 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

[8]. Bahnsen, A. C., Stojanovic, A., Aouada, D., & Ottersten, B. (2013). *Cost sensitive credit card fraud detection using Bayes minimum risk*

[9]. Barredo Arrieta, A., 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

[10]. 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

[11]. Buonaguidi, B., Mira, A., Bucheli, H., & Vitanis, V. (2022). Bayesian quickest detection of credit card fraud. *Bayesian Analysis*, *17*(1), 261–290. https://doi.org/10.1214/20-ba1254

[12]. 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

[13]. 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

[14]. Chen, T., & Guestrin, C. (2016). *XGBoost: A scalable tree boosting system* Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,

[15]. Dal Pozzolo, A., Boracchi, G., Caelen, O., Alippi, C., & Bontempi, G. (2015). *Credit card fraud detection and concept-drift adaptation with delayed supervised information*

[16]. 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

[17]. 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

[18]. Davis, J., & Goadrich, M. (2006). *The relationship between precision-recall and ROC curves*

[19]. Dornadula, V. N., & Geetha, S. (2020). *Credit card fraud detection using machine learning algorithms*

[20]. 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

[21]. 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

[22]. 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

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

[24]. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Pedreschi, D., & Giannotti, F. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys*, *51*(5), Article 93. https://doi.org/10.1145/3236009

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

[26]. Han, H., Wang, W.-Y., & Mao, B.-H. (2005). *Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning* Advances in Intelligent Computing,

[27]. Hand, D. J., Whitrow, C., Adams, N. M., Juszczak, P., & Weston, D. J. (2007). 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

[28]. 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

[29]. Jabed Hasan, T., & Waladur, R. (2023). AI-Driven Cybersecurity, IOT Networking, And Resilience Strategies For Industrial Control Systems: A Systematic Review For U.S. Critical Infrastructure Protection. *International Journal of Scientific Interdisciplinary Research*, *4*(4), 144–176. https://doi.org/10.63125/mbyhj941

[30]. Jeyaraj, A. (2020). DeLone & McLean models of information system success: Critical meta-review and research directions. *International Journal of Information Management*, *54*, 102139. https://doi.org/10.1016/j.ijinfomgt.2020.102139

[31]. Jha, S., Guillen, M., & Westland, J. C. (2012). Employing transaction aggregation strategy to detect credit card fraud. *Expert Systems with Applications*, *39*(16), 12650–12657. https://doi.org/10.1016/j.eswa.2012.05.018

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

[33]. Jullum, M., Løland, A., Huseby, R. B., Ånonsen, G., & Lorentzen, J. (2020). Detecting money laundering transactions with machine learning. *Journal of Money Laundering Control*, *23*(1), 173–186. https://doi.org/10.1108/jmlc-07-2019-0055

[34]. 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

[35]. Krivko, M. (2010). A hybrid model for plastic card fraud detection systems. *Expert Systems with Applications*, *37*(8), 6070–6076. https://doi.org/10.1016/j.eswa.2010.02.119

[36]. Kull, M., Silva Filho, T. M., & Flach, P. (2017). Beyond sigmoids: How to obtain well-calibrated probabilities from binary classifiers with beta calibration. *Electronic Journal of Statistics*, *11*(2), 5052–5080. https://doi.org/10.1214/17-ejs1338si

[37]. Lai, Y., Sun, H., & Ren, J. (2018). Understanding the determinants of big data analytics (BDA) adoption in logistics and supply chain management: An empirical investigation. *The International Journal of Logistics Management*, *29*(2), 676–703. https://doi.org/10.1108/ijlm-06-2017-0153

[38]. Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). *Isolation forest* 2008 Eighth IEEE International Conference on Data Mining,

[39]. Liu, Z., Dou, Y., Yu, P. S., Deng, Y., & Peng, H. (2020). *Alleviating the inconsistency problem of applying graph neural network to fraud detection*

[40]. Md. Akbar, H., & Sharmin, A. (2022). Neurobiotechnology-Driven Regenerative Therapy Frameworks For Post-Traumatic Neural Recovery. *American Journal of Scholarly Research and Innovation*, *1*(02), 134–170. https://doi.org/10.63125/24s6kt66

[41]. Md. Foysal, H., & Subrato, S. (2022). Data-Driven Process Optimization in Automotive Manufacturing A Machine Learning Approach To Waste Reduction And Quality Improvement. *Journal of Sustainable Development and Policy*, *1*(02), 87-133. https://doi.org/10.63125/2hk0qd38

[42]. Misra, S., Thakur, S., Ghosh, M., & Saha, S. K. (2020). *An autoencoder based model for detecting fraudulent credit card transaction*

[43]. 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

[44]. Nelson, R. R., Todd, P. A., & Wixom, B. H. (2005). Antecedents of information and system quality: An empirical examination within the context of data warehousing. *Journal of Management Information Systems*, *21*(4), 199–235. https://doi.org/10.1080/07421222.2005.11045823

[45]. 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

[46]. Olowookere, T. A., & Adewale, O. S. (2020). A framework for detecting credit card fraud with cost-sensitive meta-learning ensemble approach. *Scientific African*, *8*, e00464. https://doi.org/10.1016/j.sciaf.2020.e00464

[47]. Petter, S., DeLone, W., & McLean, E. R. (2008). Measuring information systems success: Models, dimensions, measures, and interrelationships. *European Journal of Information Systems*, *17*(3), 236–263. https://doi.org/10.1057/ejis.2008.15

[48]. Pourhabibi, T., Ong, K.-L., Kam, B., & Boo, Y. L. (2020). Fraud detection: A systematic literature review of graph-based anomaly detection approaches. *Decision Support Systems*, *133*, 113303. https://doi.org/10.1016/j.dss.2020.113303

[49]. 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

[50]. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you?: Explaining the predictions of any classifier. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations,

[51]. Rifat, C., & Rebeka, S. (2023). The Role Of ERP-Integrated Decision Support Systems In Enhancing Efficiency And Coordination In Healthcare Logistics: A Quantitative Study. *International Journal of Scientific Interdisciplinary Research*, *4*(4), 265–285. https://doi.org/10.63125/c7srk144

[52]. 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

[53]. Somasundaram, A., & Reddy, S. (2019). Parallel and incremental credit card fraud detection model to handle concept drift and data imbalance. *Neural Computing and Applications*, *31*(Suppl 1), 3–14. https://doi.org/10.1007/s00521-018-3633-8

[54]. Stieninger, M., Nedbal, D., Wetzlinger, W., Wagner, G., & Erskine, M. A. (2014). Impacts on the organizational adoption of cloud computing: A reconceptualization of influencing factors. *Procedia Technology*, *16*, 85–93. https://doi.org/10.1016/j.protcy.2014.10.071

[55]. Sun, Y., Kamel, M. S., Wong, A. K. C., & Wang, Y. (2007). Cost-sensitive boosting for classification of imbalanced data. *Pattern Recognition*, *40*(12), 3358–3378. https://doi.org/10.1016/j.patcog.2007.04.009

[56]. 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

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

[58]. Wallace, B. C., & Dahabreh, I. J. (2012). *Class probability estimates are unreliable for imbalanced data (and how to fix them)*

[59]. 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

[60]. Zhang, Y., & Trubey, P. (2019). Machine learning and sampling scheme: An empirical study of money laundering detection. *Computational Economics*, *54*(4), 1043–1063. https://doi.org/10.1007/s10614-018-9864-z

[61]. Zulqarnain, F. N. U. (2022). Policy Optimization for Sustainable Energy Security: Data-Driven Comparative Analysis Between The U.S. And South Asia. *American Journal of Interdisciplinary Studies*, *3*(04), 294-331. https://doi.org/10.63125/v4e4m413

[62]. Zulqarnain, F. N. U., & Subrato, S. (2021). Modeling Clean-Energy Governance Through Data-Intensive Computing And Smart Forecasting Systems. *International Journal of Scientific Interdisciplinary Research*, *2*(2), 128–167. https://doi.org/10.63125/wnd6qs51

[63]. Zulqarnain, F. N. U., & Subrato, S. (2023). Intelligent Climate Risk Modeling For Robust Energy Resilience And National Security. *Journal of Sustainable Development and Policy*, *2*(04), 218-256. https://doi.org/10.63125/jmer2r39