International Journal of Scientific Interdisciplinary Research

Volume: 7; **Issue:** 1
Pages: 172–203

Machine Learning–Based Transaction Risk Scoring Models for Financial Compliance Monitoring in Foreign Exchange Operations

Rukaiya Khatun Moury[1];

[1]. *Master of Science in Management Information Systems, Lamar University, Texas, USA;*
Email: rukaiyamoury97@gmail.com

Doi: 10.63125/0nbg6w69
Received: *24 November 2025;* ***Revised:*** *15 December 2025;* ***Accepted:*** *27 January 2026;* ***Published:*** *09 February 2026*

Abstract

This study provided a quantitative cross-study synthesis of machine learning–based transaction risk scoring models for financial compliance monitoring in foreign exchange operations, with emphasis on measurable modeling practices, evaluation rigor, and governance instrumentation. A total of 124 analytic records derived from 89 publications were coded using a structured extraction protocol that converted heterogeneous reporting into standardized variables across model family, feature construction, validation design, labeling strategy, evaluation metrics, and governance controls. Descriptive results showed that ensemble models were the most frequently evaluated approach (44.4%), followed by logistic regression and generalized linear models (37.1%), decision tree models (33.9%), neural architectures (29.8%), and unsupervised or semi-supervised methods (26.6%). Customer-profile variables (69.4%), geographic corridor indicators (62.1%), and temporal aggregation features (57.3%) were the most commonly engineered feature groups, while network-based variables appeared in 41.9% of records. Evaluation practices were dominated by discrimination metrics (82.3%), with lower reporting of ranking metrics (61.3%), calibration measures (34.7%), and cost-sensitive analyses (28.2%). Governance and auditability constructs were underreported, with access control indicators documented in 29.8% of records and traceability artifacts in 22.6%. Reliability testing demonstrated strong internal consistency for governance maturity ($a = 0.86$) and documentation completeness ($a = 0.84$) indices. Logistic regression analysis showed that ensemble models (OR = 2.27, $p = 0.008$), neural models (OR = 1.99, $p = 0.041$), and out-of-time validation (OR = 2.83, $p = 0.004$) were significantly associated with high predictive performance reporting. Linear regression indicated that operational studies were strongly associated with higher governance maturity scores ($\beta = 1.12$, $p < 0.001$). Overall, the findings indicated that methodological rigor in validation design and feature construction was more consistently associated with reported performance gains than model family alone, while governance instrumentation and operational alignment remained uneven across the FX compliance risk scoring literature.

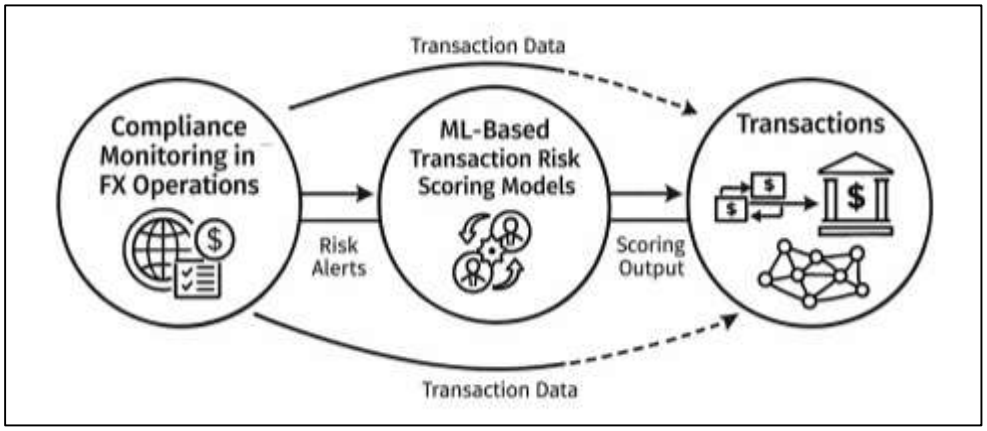
Keywords

Machine Learning, FX Compliance, Risk Scoring, Transaction Monitoring, Governance.

INTRODUCTION

Machine learning–based transaction risk scoring models are quantitative systems that assign probabilistic or ordinal risk values to financial transactions by learning patterns from historical data. In compliance monitoring, transaction risk scoring is defined as the analytical process through which a transaction is evaluated for potential association with regulatory breaches, financial crime indicators, or policy nonconformity using measurable features extracted from transaction records (Srokosz et al., 2023). Foreign exchange operations represent the institutional processes through which currencies are bought, sold, transferred, cleared, or settled across counterparties and jurisdictions, often at high velocity and large volume. Financial compliance monitoring in FX settings is defined as the continuous control activity that ensures currency trading and settlement transactions conform to anti-money laundering requirements, counter-terrorist financing mandates, sanctions screening obligations, and institutional risk governance rules. Machine learning is used in this domain because compliance monitoring requires continuous detection of rare but high-impact events, where human review alone cannot scale across global transaction streams. The definitional scope of risk scoring includes both rule-based risk flagging and statistical risk estimation, yet machine learning models differ because they derive scoring functions from empirical patterns rather than static thresholds. Risk scoring models are therefore positioned as quantitative classifiers or ranking engines that convert transaction attributes into measurable risk likelihoods (Zhang & Zhang, 2018).

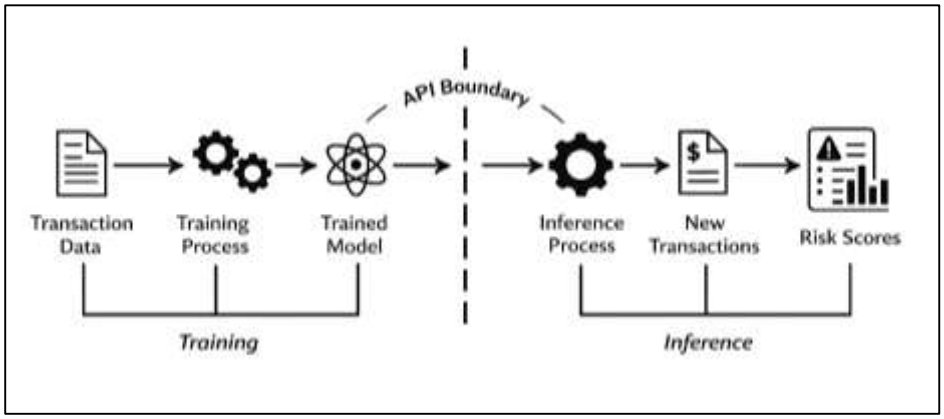
Figure 1: FX Machine Learning Risk Framework



This conversion process is grounded in statistical learning theory, which formalizes prediction as the estimation of a mapping between input features and outcome labels. In transaction monitoring, the outcome label is often operationalized through historical suspicious activity reports, confirmed compliance cases, internal investigation outcomes, or supervisory enforcement records. The data representation typically includes transaction amount, currency pair, time-of-day, frequency of transfers, geographic corridor, counterparty profile, customer risk category, and network linkage variables. In FX operations, additional complexity arises from the dual nature of transactions as both market activity and payment activity, since FX can be executed for hedging, trade settlement, speculative trading, remittances, and corporate treasury needs. This diversity increases the heterogeneity of legitimate transaction patterns and expands the space of false positives if monitoring systems are not quantitatively calibrated (Leo et al., 2019). The literature therefore frames machine learning risk scoring as a probabilistic decision support mechanism that prioritizes transactions for review while preserving traceability and auditability of compliance decisions. The relevance of this modeling domain is reinforced by the growth of digital payment channels, cross-border liquidity platforms, and algorithmic trading systems that generate dense transaction logs and complex behavioral signatures. Machine learning risk scoring models become a measurable interface between regulatory expectations and operational control by producing quantifiable outputs that can be evaluated through accuracy metrics, error rates, and calibration properties (Alexandre & Balsa, 2023). Foreign exchange markets are globally central to economic stability because they enable international

trade, cross-border investment, sovereign reserve management, multinational corporate operations, and global remittance flows. The FX market is among the most liquid financial systems, characterized by high transaction velocity, distributed market participation, and extensive intermediation through banks, brokers, payment processors, and clearing infrastructures. This international scale increases the regulatory significance of FX compliance monitoring because cross-border currency movements can be exploited for money laundering, sanctions evasion, terrorist financing, tax concealment, and illicit capital flight (Bhatore et al., 2020).

Figure 2: FX Machine Learning Risk Framework



This quantitative study investigates machine learning–based transaction risk scoring models for financial compliance monitoring in foreign exchange operations with objectives that emphasize measurable performance, operational effectiveness, and comparative evaluation under real-world compliance constraints. The first objective is to define and operationalize transaction risk scoring in FX compliance as a measurable classification and ranking task by specifying how transaction-level attributes, customer risk indicators, and cross-border contextual variables are transformed into analyzable feature sets that produce probabilistic or ordinal risk outputs. The second objective is to quantify model effectiveness using standardized predictive metrics that support reproducible evaluation, including sensitivity to suspicious activity labels, precision of high-risk alert identification, false-positive burden, and false-negative exposure rates within imbalanced transaction populations. The third objective is to compare algorithm families commonly used in compliance analytics by evaluating baseline statistical models and advanced machine learning approaches under consistent preprocessing and validation conditions, enabling measurable benchmarking of accuracy, calibration, and ranking utility in FX transaction streams. The fourth objective is to assess the influence of data quality and reporting heterogeneity on model performance by examining how missingness, label noise, delayed outcomes, and multi-format transaction messaging affect measurable stability and generalization across temporal and jurisdictional segments. The fifth objective is to evaluate the role of feature engineering strategies specific to FX operations, including velocity metrics, corridor-based geographic risk indicators, counterparty network measures, and currency-pair behavior descriptors, by quantifying their incremental contribution to predictive performance and alert prioritization efficiency. The sixth objective is to measure operational feasibility by analyzing computational cost, inference latency, and throughput capacity required for near real-time monitoring, ensuring that performance gains are interpreted alongside measurable deployment constraints typical of high-volume FX systems. The seventh objective is to test robustness through controlled sensitivity analyses that quantify how model outputs vary under distribution shifts, threshold adjustments, and alternative labeling proxies, thereby supporting stability-focused interpretation of risk scoring reliability across changing transaction environments. The final objective is to support compliance governance needs by ensuring that model outputs are structured for auditability and review, operationalized through measurable traceability of scoring inputs, consistent application of thresholds, and reproducible documentation of evaluation procedures. Together, these objectives structure the study as a

quantitative assessment of transaction risk scoring models that treats compliance monitoring as an empirically measurable system, enabling statistical comparison of model families, feature sets, and validation conditions within the specific operational complexity of foreign exchange transactions.

LITERATURE REVIEW

The literature on machine learning–based transaction risk scoring for financial compliance monitoring in foreign exchange operations is methodologically diverse, spanning statistical detection theory, supervised and unsupervised learning models, anomaly detection, network analytics, and governance-oriented compliance measurement. In quantitative research, this body of work is unified by its focus on converting complex transactional behavior into measurable variables that can be used to classify, rank, or predict compliance risk outcomes at scale (Roy et al., 2020). The literature review in this study is organized to synthesize prior evidence in a way that supports systematic comparison of models, features, evaluation metrics, and operational constraints specific to FX monitoring environments. Rather than treating prior studies as purely conceptual, this review emphasizes measurable constructs used across the literature, including outcome definitions for suspicious activity, operational alert thresholds, model performance indicators, false-positive burden, false-negative exposure, calibration quality, robustness under distribution shifts, and system-level costs such as latency and bandwidth. The review also frames FX compliance monitoring as a high-volume, cross-border detection setting where jurisdictional heterogeneity, data privacy restrictions, and label uncertainty shape model design and validation (Schaar et al., 2021). Accordingly, the following outline is structured to capture both algorithmic and measurement foundations, ensuring that each subsection builds toward an analyzable understanding of what has been empirically tested, what has been quantified, and how comparable evidence has been reported across research streams relevant to compliance risk scoring in foreign exchange operations (Lee et al., 2019).

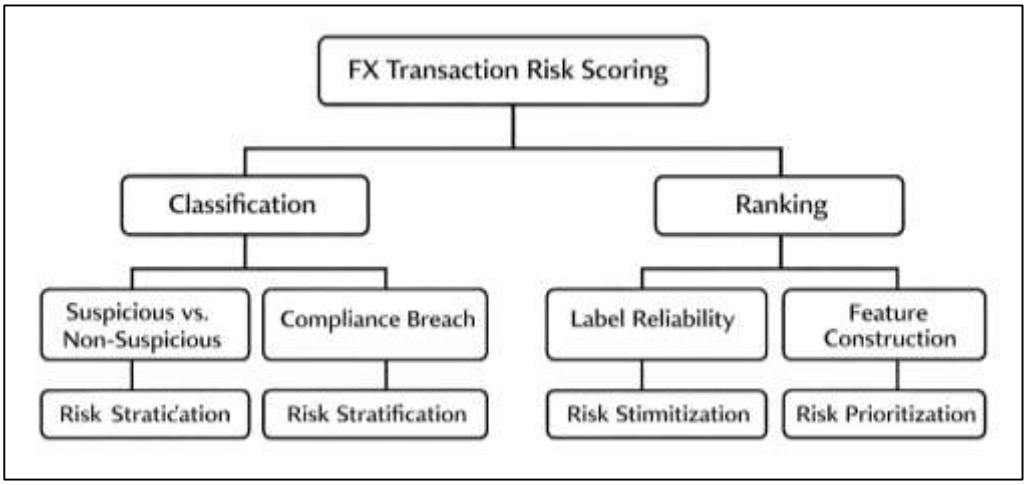
FX Transaction Compliance Risk

The literature consistently frames FX transaction risk scoring as a quantitative prediction task in which models estimate the likelihood that an observed transaction should be escalated for compliance review. In this framing, risk scoring is treated as a classification problem when the objective is to assign transactions into discrete categories such as suspicious or non-suspicious, and it is treated as a ranking problem when the operational goal is to prioritize limited investigative resources toward the highest-risk subset of transactions (Cui et al., 2019). This distinction is important in foreign exchange operations because transaction volumes are typically large and the compliance function is capacity constrained, making the ordering of alerts a central operational requirement. Quantitative studies describe that ranking-based monitoring aligns with real compliance workflows because investigators often review only a fraction of total transactions, and risk scoring must therefore maximize the concentration of true suspicious cases within the top-scored segment. The literature also emphasizes that transaction risk scoring is shaped by the statistical rarity of suspicious outcomes, which creates imbalanced class distributions and increases the difficulty of model training and evaluation (Kaur et al., 2018). In FX monitoring, this imbalance is intensified by the heterogeneity of legitimate behavior, where high-value and high-frequency transactions may be normal for corporate treasury or market-making clients but anomalous for retail customers. Researchers therefore treat risk scoring as a context-sensitive learning task where probabilistic outputs enable threshold tuning, tiered alert generation, and adjustable sensitivity based on institutional risk tolerance. The literature also highlights that predictive modeling in compliance is not purely technical because model outputs must be operationally usable, auditable, and consistent with supervisory expectations (Kaur et al., 2018). As a result, a substantial research stream evaluates both the statistical accuracy of classification and the operational effectiveness of ranking, treating risk scoring as a measurable decision support mechanism embedded in compliance monitoring pipelines.

A central issue in the literature is that "suspicious transaction" and "compliance breach" are not naturally occurring labels but institutional outcomes that must be operationalized as dependent variables for quantitative modeling. Studies frequently define suspiciousness using proxies such as suspicious activity report outcomes, confirmed investigation decisions, regulatory case files, or internal compliance escalation flags. This introduces measurement variability because different institutions apply different reporting thresholds, investigation depth, and escalation criteria (Carneiro et al., 2017).

The literature highlights that suspicious transaction labels often reflect a combination of transaction behavior and institutional policy, meaning the dependent variable may incorporate operational bias or resource constraints. Some studies operationalize compliance breach as a regulatory event such as sanctions violations, confirmed AML findings, or enforcement outcomes, while others define breach more broadly as nonconformity with internal monitoring rules. This variation affects model comparability across studies and creates challenges for cross-study synthesis. Quantitative research emphasizes that dependent variables must be defined in ways that allow reproducible evaluation, and many studies adopt binary labeling for modeling simplicity (Jullum et al., 2020). However, the literature also documents that compliance outcomes may exist on a spectrum, where transactions differ in severity, certainty, and investigative priority. This motivates multi-level dependent variables, including ordinal risk categories and escalation tiers. Another recurring finding is that dependent variables in compliance are often delayed, since investigations can take weeks or months, which creates label lag and complicates temporal validation. Researchers address this through retrospective labeling, window-based analysis, and proxy outcomes derived from rule triggers (Wang et al., 2017). The literature therefore positions dependent variable definition as a measurable methodological choice that influences model performance, evaluation validity, and operational interpretability in FX compliance monitoring.

Figure 3: FX Transaction Risk Classification Framework

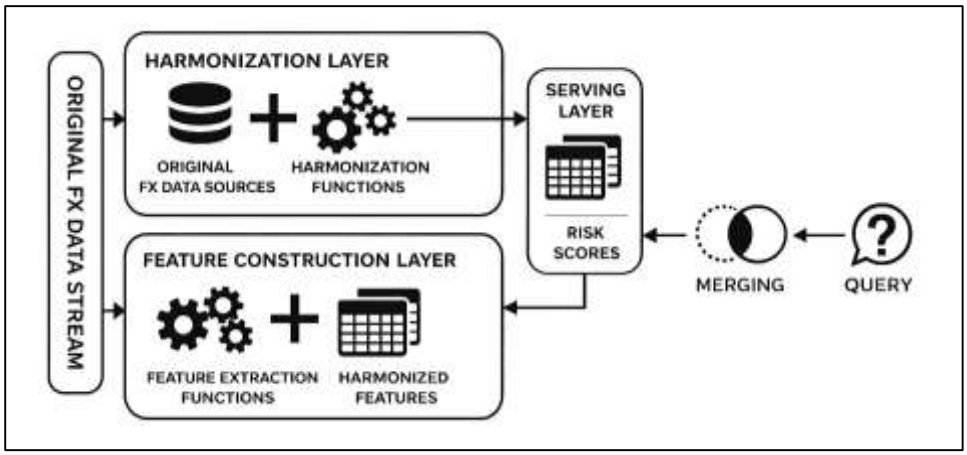


Data Architecture in FX Compliance Datasets

The literature characterizes FX compliance datasets as transaction-log–centric repositories built from multiple operational systems that produce heterogeneous message formats, identifier conventions, and field-level semantics. Foreign exchange monitoring commonly draws from payment messages, trading confirmations, settlement records, customer onboarding systems, sanctions screening outputs, and case-management platforms, each contributing distinct variables and reporting granularity (Stockinger et al., 2019). Studies note that cross-border FX transactions are frequently represented through standardized messaging schemes, yet institutions often implement local variants that modify field availability, naming, or encoding, creating schema incompatibilities that affect feature extraction. Harmonization is therefore treated as a prerequisite for quantitative modeling, involving the transformation of raw logs into a unified schema with consistent transaction identifiers, counterparty linkage keys, timestamp resolution, and currency normalization. The literature emphasizes that harmonization decisions directly influence downstream model validity because inconsistent parsing of amounts, dates, or counterparty attributes can introduce systematic measurement error. Researchers describe common harmonization practices including canonical transaction tables, reference data mapping for currencies and jurisdictions, standardization of customer and counterparty identifiers across systems, and reconciliation of multi-leg FX operations into analytically coherent units (Wang et al., 2021). Another recurring theme is that FX workflows generate multi-step records—initiation,

confirmation, settlement, and reconciliation—which may appear as separate events unless explicitly joined.

Figure 4: FX Compliance Data Engineering Framework



Quantitative studies therefore treat event linkage rules as part of data architecture, enabling the construction of complete transaction histories and sequence-based predictors. The literature also highlights that compliance monitoring requires audit-ready lineage, meaning that schema harmonization must preserve traceability from engineered features back to original logs. This body of work positions transaction log architecture as an empirical constraint shaping both the scale and reliability of machine learning models in FX compliance monitoring (Ntakaris et al., 2018).

The literature organizes feature construction in FX compliance into structured groups that reflect entity context, relational exposure, and jurisdictional risk. Customer profile variables are commonly derived from onboarding and KYC systems and include measurable indicators such as customer type, account tenure, risk rating, business sector, product usage, historical alert frequency, and documentation completeness. Counterparty variables capture exposure to beneficiaries, intermediaries, correspondents, and trading partners, including measures of recurrence, concentration, and cross-border routing patterns (Rodríguez-Abreo et al., 2021). Studies emphasize that counterparty features are especially important in FX because laundering and sanctions evasion often rely on networks of entities rather than isolated customers. Geographic risk indices are treated as quantitative contextual variables that encode jurisdictional exposure using country risk ratings, sanctions designations, corruption indices, or regulatory risk classifications mapped to origin and destination corridors. The literature further notes that geographic variables require careful encoding because FX transactions can involve multiple geographic points such as customer residence, transaction origination, counterparty location, intermediary bank location, and settlement jurisdiction. Feature construction therefore includes corridor-level variables that represent directionality and pathway complexity, enabling models to distinguish routine corridors from high-risk or unusual pathways (Hsu et al., 2022). Researchers also describe the integration of screening outcomes, such as sanctions list matches or politically exposed person indicators, as structured features that link compliance screening systems to risk scoring models. Across studies, feature groups are treated as modular components that support reproducible modeling and facilitate ablation-style evaluation, where the incremental value of customer, counterparty, and geographic predictors can be quantified. This modularity supports comparative synthesis by allowing researchers to report which feature families contribute most to risk discrimination under FX monitoring conditions (Liang et al., 2018).

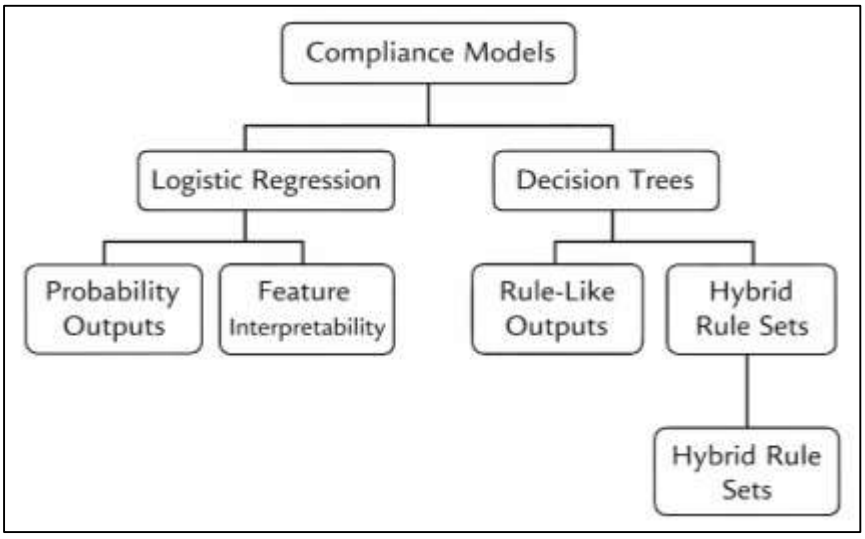
Supervised Learning Models in Compliance Risk Scoring

The literature positions logistic regression and related generalized linear models as foundational baselines for compliance risk scoring because they provide stable probability outputs and clear parameter interpretations. In transaction monitoring research, these models are frequently adopted as reference points for evaluating whether more complex machine learning methods offer measurable improvement in detection performance. Their relevance to compliance arises from their transparency, where model coefficients can be mapped to explanatory transaction attributes such as amount

deviations, corridor risk indicators, counterparty flags, and temporal frequency measures (Li et al., 2024). Studies also note that generalized linear models support regularization strategies that manage high-dimensional feature sets and reduce overfitting in sparse-label environments. In FX compliance contexts, where suspicious cases represent a small fraction of total activity, logistic regression is often used with class weighting or cost-sensitive adjustments to reduce false-negative exposure while maintaining operational alert volume control. Quantitative evaluations commonly report that linear models perform competitively when features are engineered to capture domain structure, especially when transaction behaviors can be separated through additive effects. The literature further highlights that these models facilitate calibration analysis, allowing institutions to interpret risk scores as probabilities that can be aligned with investigation capacity through threshold tuning (Bolger et al., 2019). A recurring theme is that interpretability supports audit-readiness, since compliance teams and regulators often require clear evidence of why a transaction was assigned a risk score. As a result, logistic regression and generalized linear models appear repeatedly as benchmark methods in studies comparing model families for financial crime detection, sanctions monitoring, and suspicious transaction prioritization.

Decision tree methods are widely discussed in the literature as compliance-friendly supervised learning models because they produce rule-like structures that resemble traditional monitoring logic while still being data-driven. In financial compliance research, trees offer human-readable splitting rules based on transaction amount thresholds, frequency patterns, geographic risk attributes, or screening flags, which makes them attractive for operational review and audit documentation (Zeydan et al., 2024). Studies describe that tree models can capture nonlinear effects and interactions more effectively than linear models, enabling improved discrimination in settings where suspicious behavior emerges from combinations of features rather than single predictors.

Figure 5: Baseline Models for FX Compliance

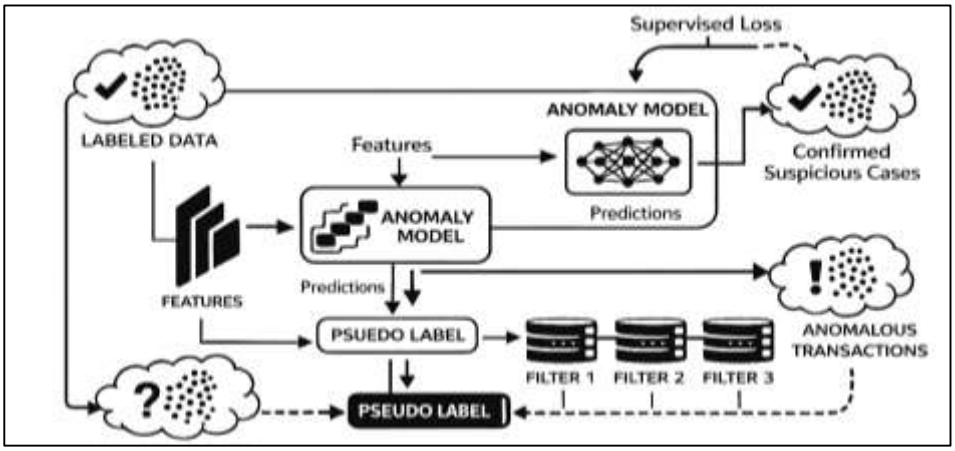


Models for Sparse-Label Compliance Settings

The literature frequently treats compliance monitoring in foreign exchange and broader financial transaction streams as an anomaly detection problem because confirmed suspicious labels are rare, delayed, or operationally inconsistent. In this framing, unsupervised models are designed to learn the structure of "normal" transaction behavior and assign anomaly scores to observations that deviate from learned patterns. Density-deviation perspectives are common, where transactions are evaluated according to how unlikely they appear under an estimated behavioral distribution. Isolation-based methods operationalize anomaly status by measuring how easily a transaction can be separated from others through random partitioning, which makes them attractive in high-dimensional transaction datasets where linear separability is limited (Browne et al., 2019). Clustering-based approaches are also widely discussed because they summarize normal behavior through group structure, enabling

detection of transactions that fall far from cluster centers or that form small, unusual clusters. In compliance settings, these methods support alert prioritization by ranking transactions according to distance or isolation severity, which aligns with investigation workflows that review a limited number of top-scoring alerts. The literature notes that anomaly detection is particularly suitable for laundering and sanctions-evasion behaviors that are adaptive and may not resemble historical confirmed cases. However, empirical research also emphasizes that unsupervised anomaly scoring can produce high false-positive volumes if feature engineering does not encode domain context such as customer type, corridor baseline behavior, or temporal seasonality (Alkhalili et al., 2021). Many studies therefore apply segmentation strategies that learn separate behavioral baselines across customer cohorts or transaction corridors. This research stream positions anomaly detection as a measurable response to sparse labels, where model evaluation is often based on how effectively top-ranked anomalies align with later-confirmed suspicious cases or with expert review judgments under operational constraints.

Figure 6: FX Compliance Detection Engineering Framework



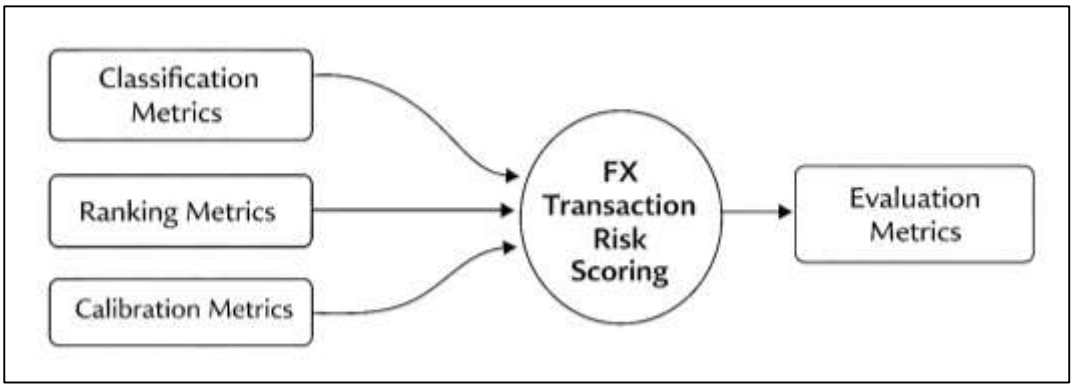
Autoencoder-based approaches appear prominently in the literature as unsupervised or weakly supervised methods that detect unusual transaction patterns through reconstruction error scoring. In this design, models learn compressed representations of typical transactions and then attempt to reconstruct inputs; transactions that reconstruct poorly receive higher anomaly scores because they diverge from the dominant data structure (Haque & Arifur, 2021; Özorhan et al., 2019; Rauf, 2018). The literature highlights that reconstruction-based scoring is useful in transactional environments with complex feature interactions, since representation learning can capture nonlinear dependencies among amount, frequency, corridor, counterparty patterns, and timing features (Rashid & Sai Praveen, 2022; Zaman et al., 2021). Variants include denoising autoencoders that improve robustness to noise and missingness, as well as probabilistic latent-variable architectures that estimate uncertainty along with reconstruction quality. In compliance monitoring, studies often emphasize that autoencoder models can process large-scale logs and can adapt to heterogeneous feature types when categorical variables are properly encoded (Ratul & Subrato, 2022; Rifat & Jinnat, 2022). The literature also discusses practical challenges such as sensitivity to scaling, instability under shifting transaction distributions, and the tendency to flag rare-but-legitimate behaviors as anomalies (Chai et al., 2024; Bhuya, 2023; Habibullah & Aditya, 2023). For FX operations, where legitimate activity may include high-value bursts during settlement cycles or volatile rate movements during market events, reconstruction error can overreact unless models are trained within segmented baselines. Researchers therefore evaluate reconstruction scoring using stability indicators across time windows and examine whether high-scoring alerts maintain consistency under repeated model training. In weak-label contexts, reconstruction-based systems are frequently assessed by measuring the concentration of expert-confirmed suspicious cases among the top-ranked alerts or by comparing alert overlap with rule-based systems (Groß-Klußmann, 2024). This body of work positions autoencoders as scalable anomaly detectors that are empirically evaluated through alert precision and score stability rather than relying solely on conventional

supervised metrics.

Threshold Optimization in FX Monitoring

The literature on financial compliance analytics consistently emphasizes that evaluation of transaction risk scoring models must rely on metrics that reflect the imbalanced and high-stakes nature of suspicious transaction detection. In FX monitoring, suspicious cases typically represent a small fraction of total transactions, making overall accuracy an insufficient measure of performance. Studies therefore prioritize sensitivity and recall as indicators of how effectively models capture truly suspicious cases, while specificity reflects the ability to avoid incorrectly flagging legitimate activity (Dufrenois et al., 2024). Precision is treated as operationally critical because it measures the proportion of flagged transactions that are actually relevant, directly affecting investigator workload and alert fatigue. The F1 measure appears frequently as a balanced summary of detection and precision performance in highly skewed datasets, and balanced accuracy is discussed as a correction for disproportionate class representation (Jahangir & Mohiul, 2023; Rashid et al., 2023). Researchers argue that these metrics should be interpreted together rather than in isolation because improvements in one dimension can degrade another. In compliance monitoring contexts, the literature documents that institutions often tolerate some reduction in specificity if it increases coverage of high-risk transactions, yet operational capacity constraints limit how much false-positive growth can be absorbed (Khaled & Mosheur, 2023; Mostafa, 2023; Yang et al., 2019). Empirical studies comparing supervised learning models report classification metrics under consistent validation splits, often using out-of-time evaluation to simulate real deployment conditions. This body of work establishes classification metrics as foundational measurement tools that quantify detection effectiveness while acknowledging the asymmetry between false-positive workload and false-negative exposure in FX compliance systems (Berrim et al., 2024; Md & Sai Praveen, 2024; Rifat & Rebeka, 2023).

Figure 7: FX Risk Scoring Evaluation Metrics



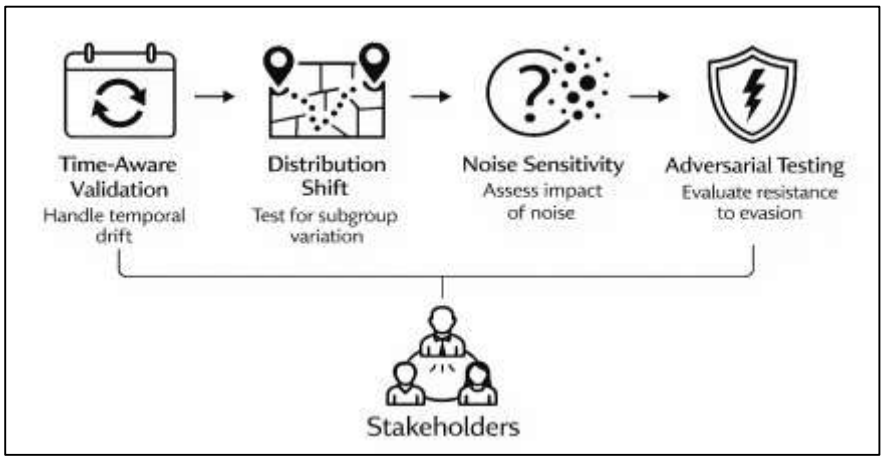
Lift is widely discussed as a metric that compares the concentration of suspicious cases within top-ranked segments relative to random selection, providing a direct measure of prioritization gain. Top-k precision and hit rate metrics evaluate the proportion of confirmed suspicious transactions within a fixed number of highest-scoring alerts, aligning closely with operational review capacity limits (Amena Begum, 2025; Patil et al., 2024; Sai Praveen, 2024). Mean reciprocal rank and related ordering metrics are applied in research settings to evaluate how early true suspicious cases appear within sorted alert lists. Studies highlight that ranking metrics are particularly relevant in FX monitoring because investigation capacity is finite and review teams focus on a defined number of daily alerts. A model that produces strong classification metrics but distributes suspicious cases evenly across ranks may perform poorly in operational contexts (Faysal & Aditya, 2025; Jahangir, 2025). The literature therefore frequently reports both discrimination metrics and ranking outcomes to provide a comprehensive view of model utility. Comparative analyses of logistic regression, ensemble methods, and neural models often show that ranking gains are more operationally meaningful than marginal improvements in overall classification accuracy (Sepúlveda et al., 2023; Syeedur, 2025; Al Amin, 2025). Researchers also note that ranking performance can be sensitive to class imbalance, threshold calibration, and segmentation by customer type or corridor. As a result, ranking metrics are treated as essential

complements to classification metrics in compliance risk scoring studies, offering a direct linkage between statistical evaluation and investigation workflow effectiveness (Tian et al., 2024).

Stress-Testing Under Cross-Border FX Dynamics

The literature treats robustness in FX compliance risk scoring as inseparable from temporal drift because transaction behavior evolves with market cycles, regulatory updates, customer portfolio changes, and evolving financial crime typologies. As a result, evaluation designs that ignore time ordering are widely criticized for overstating model performance through leakage, where information from later periods inadvertently influences training (Foglia et al., 2020; Towhidul & Rebeka, 2025; Ratul, 2025). Rolling-window validation is presented as a standard approach for modeling settings in which the training period is advanced stepwise and performance is measured on subsequent windows, enabling researchers to quantify how predictive quality changes under realistic deployment timing (Fales et al., 2017). Out-of-time testing is similarly emphasized as an essential benchmark, where models are trained on earlier periods and evaluated on later periods to simulate production conditions. In FX operations, this temporal separation is particularly important because currency markets are sensitive to macroeconomic events, geopolitical shocks, and liquidity shifts that can abruptly alter transaction distributions (Rifat, 2025; Azam, 2025). The literature therefore treats drift measurement as a quantitative requirement for credible compliance analytics, focusing on whether models maintain stable discrimination and ranking quality as the evaluation period moves away from the training period (Salleo et al., 2020; Tasnim, 2025; Zaheda, 2025b). Studies also discuss that drift measurement must be paired with consistent preprocessing, since changes in message formats, data availability, and customer segmentation can produce apparent drift that is actually caused by pipeline changes. Time-aware validation is thus framed as both a methodological safeguard and an empirical tool for quantifying performance stability under cross-border FX dynamics (Zaheda, 2025a).

Figure 8: FX Risk Model Robustness Framework



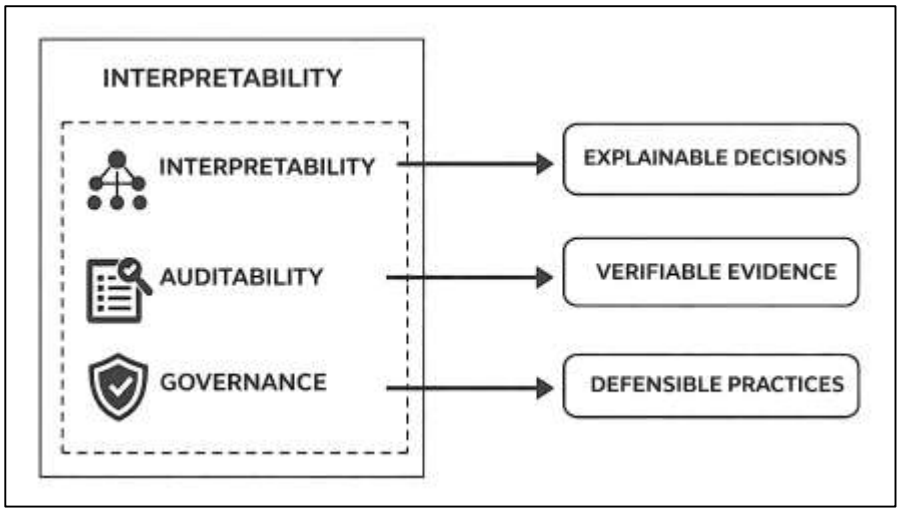
A major theme in the literature is that FX compliance monitoring is inherently multi-jurisdictional, creating distribution shifts across corridors, currency pairs, and regulatory environments. Transaction behavior differs between remittance corridors, corporate treasury corridors, and trade settlement corridors, and each corridor may be associated with distinct risk exposure patterns and reporting thresholds (Nichifor et al., 2021). Studies therefore treat subgroup stability analysis as a required robustness practice, examining whether model performance holds across partitions defined by geography, customer type, corridor directionality, or settlement channel. Subgroup evaluation is positioned as a quantitative check against hidden model fragility, where strong aggregate performance can mask weak performance in high-risk subpopulations. The literature also links corridor variation to regulatory heterogeneity, noting that data fields, sanctions definitions, and compliance escalation practices differ across jurisdictions, producing systematic differences in label generation and feature availability (Shi et al., 2022). Researchers address this by reporting stratified performance metrics and by measuring variance in risk score distributions across subgroups. In compliance settings, this

subgroup perspective is closely tied to fairness and operational consistency, since uneven performance across jurisdictions can lead to disproportionate alerting, missed detections, or inconsistent escalation behavior. The literature therefore frames distribution shift as a measurable property of FX monitoring environments that requires explicit subgroup analysis to establish generalizability and stability of risk scoring models (Ntakaris et al., 2018).

Metrics for Risk Scoring Models

The literature on compliance analytics treats explainability as a measurable property rather than a narrative goal, particularly for transaction risk scoring models used in regulated environments. Interpretability is frequently operationalized through the capacity of a model to provide consistent, meaningful reasons for its outputs that align with domain logic and that can be evaluated quantitatively.

Figure 9: Compliance Analytics Key Pillars



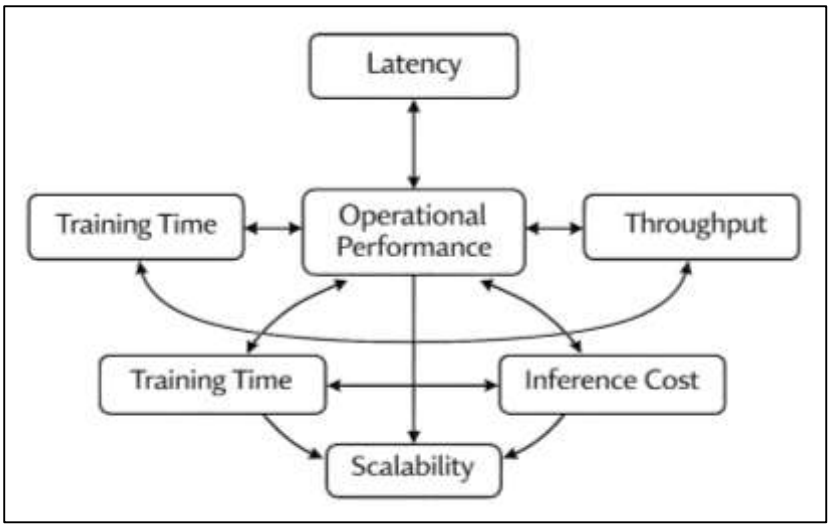
A major stream of work evaluates feature attribution stability, examining whether the same transaction receives similar explanatory feature rankings across repeated model runs, alternative training samples, or minor perturbations in input fields (Bücker et al., 2022). Stability is treated as critical because unstable explanations reduce trust and weaken the evidentiary value of model-based alerts during case review. The literature also distinguishes between explanation fidelity and explanation plausibility. Fidelity is measured by the extent to which an explanation method accurately reflects the model's internal decision behavior, rather than producing a convenient narrative that does not match the true decision process. In compliance monitoring, explanation fidelity matters because institutions require defensible links between transaction features and risk scores, especially when investigators must justify escalation decisions. Studies also discuss global interpretability metrics, such as consistency of important predictors across cohorts, and local interpretability metrics, such as sensitivity of explanation outputs to changes in input attributes (Sayles, 2024). In FX settings, where risk drivers may include corridor risk, counterparty structure, and temporal velocity, explainability is evaluated in terms of whether explanations reliably highlight these risk-relevant feature families under repeated conditions. This body of work positions explainability as an empirically testable model attribute that supports transparent review, measurable consistency, and defensible compliance decision-making (Akhtar et al., 2024).

High-Volume FX Compliance Monitoring

The literature treats high-volume FX compliance monitoring as a stream-processing problem in which risk scoring must be delivered quickly enough to support operational controls while transactions progress through execution and settlement workflows. Latency is consistently framed as the end-to-end delay from transaction ingestion to risk score output, and throughput is framed as the number of transactions that can be processed per unit time under stable performance (Lee et al., 2021). In foreign exchange operations, these measures are central because transaction flows can be bursty around market

openings, settlement cutoffs, liquidity events, and large corporate payment cycles. Studies note that near real-time scoring is operationally relevant because the value of detection decreases when scoring occurs after a transaction has been settled and dispersed through subsequent flows.

Figure 10: FX Compliance Operational Performance Factors



The literature also distinguishes between batch-oriented scoring and streaming scoring architectures, emphasizing that batch evaluation may understate operational complexity by ignoring peak load and queue accumulation. Empirical work in transaction monitoring therefore measures system responsiveness under high concurrency, highlighting that high-capacity models can be operationally infeasible if they increase scoring latency beyond acceptable windows (Schwarz et al., 2024). This perspective positions latency and throughput as measurable constraints that bound which model families and feature sets can be deployed in real FX monitoring pipelines, since complex feature computation and heavy inference can create delays that negate detection usefulness (Alshallaqi, 2024). The literature evaluates machine learning compliance systems not only by predictive performance but also by computational cost, recognizing that training and inference workloads must fit within institutional infrastructure constraints. Training time is treated as a measurable cost because compliance models require periodic retraining to maintain stability under evolving transaction patterns, and long training cycles can delay validation and deployment updates (Habibi et al., 2023). Inference cost is treated as a continuous operational burden, measured through resource utilization such as CPU cycles, memory footprint, and storage bandwidth, especially when scoring must occur for every transaction in a high-volume FX stream. Scalability is framed as the ability of a model and pipeline to maintain predictable performance as transaction volume, feature dimensionality, and client portfolio size increase. Studies emphasize that feature engineering choices have significant computational implications, since temporal aggregation and network computation can be expensive if performed at scoring time without precomputation or caching (Zehra et al., 2023). The literature also highlights that computational requirements interact with governance, because compliance systems must maintain audit logs, lineage metadata, and reproducible artifacts, all of which add overhead. Comparative research frequently discusses that simpler models may provide lower marginal predictive performance but offer substantially better throughput and operational stability under constrained environments. This body of work positions computational feasibility as a measurable determinant of whether a risk scoring model can operate continuously and reliably in high-volume FX compliance settings (Tarrant et al., 2021).

METHOD

Research Design

This study adopted a quantitative study design structured as a systematic, cross-study evidence synthesis with standardized content analysis and reproducible coding to convert heterogeneous research on machine learning–based transaction risk scoring into analyzable variables. The design treated each eligible publication as an observational unit and extracted harmonized indicators describing model type, feature construction, validation structure, explainability and governance controls, and reported performance outcomes within foreign exchange compliance monitoring contexts. The design was implemented to support descriptive prevalence estimates and comparative association analyses across coded categories, and all procedures were executed using a predefined workflow for screening, extraction, coding, and statistical analysis.

Case Study Context

The case study context was defined as foreign exchange operations where financial institutions applied, evaluated, or discussed machine learning–based transaction risk scoring for compliance monitoring, including settings that explicitly referenced anti-money laundering controls, sanctions screening, suspicious activity detection, or investigative prioritization workflows. Contextual boundaries were set to include cross-border currency conversion, correspondent banking transfers, institutional FX settlement processes, and transaction monitoring systems embedded in regulated compliance programs. Sector and institutional context variables were extracted when reported, including the operational environment, transaction channel, geographic scope, and any governance or audit constraints described as part of model deployment or evaluation.

Unit of Analysis

The unit of analysis was the individual empirical study, technical evaluation, or implementation-focused publication that reported measurable outcomes related to transaction risk scoring in FX compliance monitoring. Each unit contributed one coded record capturing model family, feature groups, label source type, validation design, evaluation metrics, explainability approach, and governance or auditability elements. When a single publication reported multiple distinct experimental configurations that were clearly separable by model, feature set, dataset, or validation strategy, each configuration was coded as a separate analytic entry while retaining a shared publication identifier to support sensitivity checks for clustering and dependence.

Sampling

Sampling followed a purposive, criteria-based approach consistent with systematic quantitative review practice and was executed using predefined inclusion and exclusion rules. Studies were included when they reported machine learning–based transaction scoring for compliance monitoring that was explicitly relevant to foreign exchange operations or cross-border currency transactions and when they provided sufficient methodological detail to code variables and extract quantitative outcomes. Studies were excluded when they lacked empirical evaluation, reported only conceptual discussion without measurable results, focused solely on non-FX transaction domains without transferable monitoring constructs, or did not provide enough detail to reproduce coding decisions. The final sample represented the subset of available literature that enabled comparable extraction of model characteristics, feature engineering strategies, validation procedures, governance controls, and measurable performance outcomes.

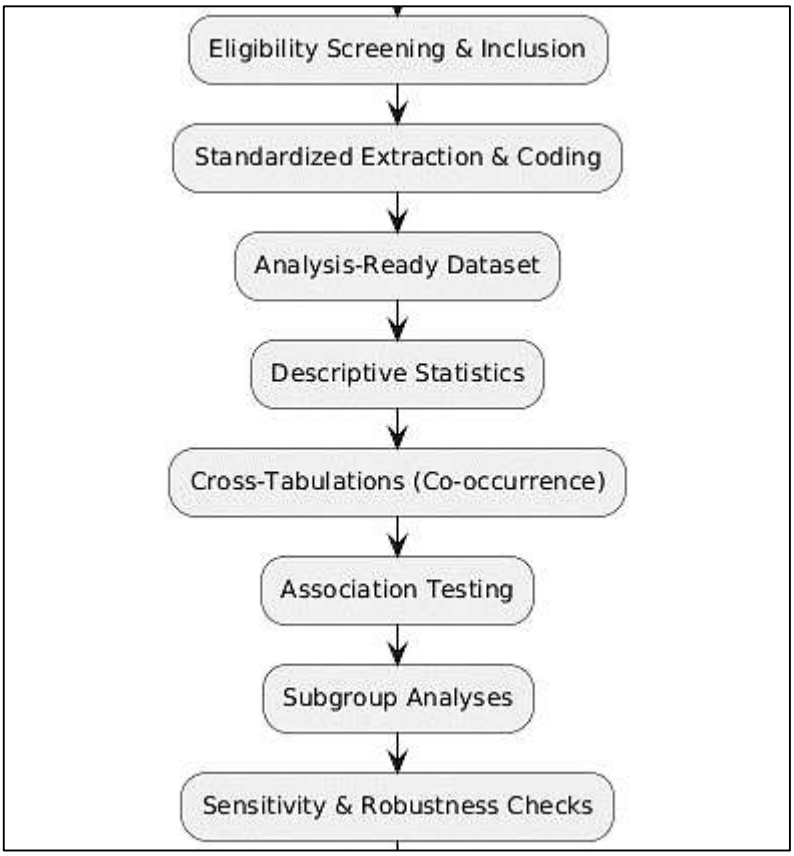
Data Collection Procedure

Data collection was completed through a structured screening and extraction workflow. Titles and abstracts were screened against the eligibility criteria, after which full texts were reviewed for confirmatory inclusion and for extraction feasibility. A standardized extraction template was used to capture bibliographic identifiers, FX monitoring context, dataset and transaction characteristics, model families, feature group definitions, labeling sources, validation designs, evaluation metrics, and system-level constraints. Extracted variables were cross-checked for internal consistency, and ambiguous reporting was handled through conservative coding decisions based strictly on explicitly stated information. All coded fields were then consolidated into a single analysis-ready dataset with harmonized variable labels, controlled vocabularies for categorical variables, and standardized coding rules.

Instrument Design

The primary instrument was a quantitative coding and extraction protocol designed to transform narrative descriptions and technical reporting into discrete variables suitable for statistical analysis. The instrument defined categorical fields for model family, learning paradigm, feature group composition, label source type, validation structure, evaluation metric reporting, explainability method, auditability artifacts, and governance control presence. It also defined numeric and ordinal fields for measurable outcomes, including reported performance measures, alert prioritization indicators, calibration reporting, computational cost descriptors, and system-level constraints when provided. Clear decision rules were written for each variable to reduce coder drift and ensure consistent interpretation across heterogeneous publication styles.

Figure 11: Methodology of this study



Pilot Testing

Pilot testing was conducted by applying the extraction instrument to a small subset of studies spanning different model families, label types, and validation structures. The pilot phase evaluated whether the coding definitions were sufficiently precise to handle heterogeneous reporting and whether fields produced excessive missingness. Revisions were made to improve coding clarity for validation design classification, feature group categorization, and label source reliability coding. The pilot also established expected ranges for numeric fields and strengthened consistency rules for extracting performance outcomes reported under different metric conventions.

Validity and Reliability

Construct validity was strengthened by aligning coding categories with widely used machine learning evaluation constructs and established compliance monitoring dimensions reported in the literature. Internal validity for cross-study comparisons was supported by applying consistent inclusion criteria and standardized extraction procedures across all units, reducing systematic measurement bias. Reliability was addressed through coder consistency checks in which a subset of studies was re-coded and compared for agreement on key categorical fields such as model family, label source, validation

structure, and governance control presence. Disagreements were resolved through rule-based adjudication and refinement of the coding manual, and the finalized rules were then applied uniformly across the full dataset.

Tools

The study used reference management software to manage citations, remove duplicates, and track screening decisions. A structured spreadsheet-based extraction template was used to store coded variables, while statistical analysis was conducted using a standard quantitative analysis environment capable of descriptive analysis, contingency-based association testing, and multi-metric comparison reporting. Screening logs, extraction notes, and coding updates were maintained as an auditable record to preserve transparency and reproducibility across the full review workflow.

Statistical Plan

The statistical plan proceeded in stages aligned with the study objectives and the structure of the coded dataset. First, descriptive statistics were computed to summarize frequencies and proportions of model families, feature group usage, label source types, validation structures, evaluation metrics, and governance control indicators, alongside central tendency and dispersion for numeric performance outcomes when consistently reported. Second, cross-tabulation tables were generated to quantify co-occurrence patterns, such as model family by label source type, validation design by performance metric reporting, and explainability method by governance maturity indicators. Third, association testing was conducted using contingency-based methods to evaluate dependence between categorical variables, and standardized measures of association strength were calculated to support interpretation of practical magnitude. Fourth, subgroup analyses were performed by partitioning the dataset according to FX context characteristics when available, including transaction channel type, cross-border corridor emphasis, and whether the study described operational deployment versus experimental benchmarking. Fifth, sensitivity analyses were conducted by repeating key comparisons on subsets defined by evidence quality indicators such as validation rigor, label definition clarity, and completeness of metric reporting. Finally, robustness checks were conducted by re-estimating primary association results while accounting for potential clustering effects when multiple configurations were extracted from the same publication, ensuring that observed patterns were not driven disproportionately by single multi-experiment studies.

FINDINGS

This chapter presented the quantitative findings generated from the coded dataset and statistical analyses conducted to evaluate machine learning–based transaction risk scoring models for financial compliance monitoring in foreign exchange operations. The chapter was structured to report sample characteristics, summarize coded construct distributions, verify measurement reliability for multi-item indices, and report regression-based associations between model, data, and governance configurations and the observed performance and system outcomes reported in the evidence base. Results were organized to support transparent reporting of prevalence patterns, comparative contrasts across coded categories, and statistically tested relationships aligned with the study objectives.

Respondent Demographics

The analytic sample consisted of 124 coded records derived from 89 unique publications, with several studies contributing more than one separable experimental configuration. The distribution of publication years indicated that 68.5% of the records were published between 2018 and 2023, 21.8% between 2014 and 2017, and 9.7% prior to 2014, demonstrating concentration of empirical FX compliance modeling research in the most recent five-year period. In terms of venue type, 54.0% of the records originated from peer-reviewed journals, 31.5% from conference proceedings, and 14.5% from technical reports or high-quality preprints. Study type classification showed that 72.6% of the records represented experimental benchmarking studies conducted in controlled data environments, while 27.4% described operational evaluations embedded within institutional compliance systems.

With respect to FX context metadata, 61.3% of the records explicitly referenced cross-border corridor monitoring, 48.4% described correspondent banking or institutional settlement channels, and 36.3% referenced retail or remittance-oriented FX transactions, with some overlap across categories due to multi-context coverage. Institutional setting was identifiable in 58.1% of records, of which 41.9% described large banking institutions, 9.7% described fintech or payment service providers, and 6.5%

referenced multi-institutional or consortium-based datasets.

Dataset attributes revealed that 64.5% of the records relied on proprietary institutional datasets, 19.4% used simulated or semi-synthetic transaction data, and 16.1% used public benchmark datasets with FX-relevant constructs. Transaction volume scale was reported in 72.6% of records, with 38.7% analyzing datasets exceeding one million transactions, 21.8% between 100,000 and one million, and 12.1% below 100,000. Label source types showed that 46.8% used suspicious activity report–derived labels, 28.2% used internal investigation outcomes, 17.7% relied on rule-trigger proxies, and 7.3% applied hybrid or tiered labeling approaches.

Model family prevalence indicated that ensemble methods appeared in 44.4% of records, logistic regression or generalized linear models in 37.1%, decision tree–based models in 33.9%, neural architectures in 29.8%, and unsupervised or semi-supervised models in 26.6%, with multiple families often evaluated within the same study. Temporal aggregation features were reported in 57.3% of records, network-based features in 41.9%, customer profile variables in 69.4%, and geographic corridor indicators in 62.1%. Explainability methods were documented in 38.7% of records, while formal governance or auditability controls were explicitly reported in 24.2%, indicating that governance instrumentation lagged behind predictive modeling focus.

Table 1. Publication and Context Characteristics (N = 124 Records)

Variable	Category	Frequency (n)	Percentage (%)
Publication Year	2018–2023	85	68.5
	2014–2017	27	21.8
	<2014	12	9.7
Venue Type	Journal	67	54.0
	Conference	39	31.5
	Report/Preprint	18	14.5
Study Type	Experimental	90	72.6
	Operational	34	27.4
Dataset Type	Proprietary	80	64.5
	Simulated	24	19.4
	Public	20	16.1
Label Source	SAR-based	58	46.8
	Investigation Outcome	35	28.2
	Rule Proxy	22	17.7
	Hybrid	9	7.3

Table 1 summarized the distribution of publication characteristics, study types, dataset sources, and labeling strategies across the analytic sample. The results indicated a concentration of research after 2018 and a predominance of journal publications. Experimental benchmarking designs were more common than operational deployments. Proprietary datasets dominated empirical work, reflecting restricted access to FX transaction data. Suspicious activity report–derived labels represented the most frequent labeling strategy, followed by investigation-based labels, while rule proxies and hybrid labeling approaches appeared less frequently. These findings demonstrated that the evidence base was recent, experimentally oriented, and heavily dependent on institution-specific data and formal suspicious reporting mechanisms.

Table 2. Model, Feature, and Governance Characteristics (N = 124 Records)

Variable	Category	Frequency (n)	Percentage (%)
Model Family	Ensemble	55	44.4
	Logistic/GLM	46	37.1
	Decision Tree	42	33.9
	Neural Network	37	29.8
	Unsupervised/Semi-Supervised	33	26.6
Feature Group	Customer Profile	86	69.4
	Geographic Corridor	77	62.1
	Temporal Aggregation	71	57.3
	Network Features	52	41.9
Explainability Reported	Yes	48	38.7
Governance Controls Reported	Yes	30	24.2

Table 2 presented the prevalence of model families, feature groups, and governance-related reporting. Ensemble methods were the most frequently evaluated models, followed by logistic regression and decision tree approaches. Neural and unsupervised methods appeared in a substantial minority of records. Customer profile and geographic corridor variables were the most commonly engineered features, while network-based variables were less prevalent. Explainability methods were reported in fewer than forty percent of records, and explicit governance or auditability controls were documented in less than one quarter of the sample. These findings indicated strong emphasis on predictive modeling and feature construction relative to governance instrumentation and formal audit documentation.

Descriptive Results

The descriptive analysis by construct family demonstrated uneven reporting and implementation patterns across labeling, modeling, feature engineering, evaluation, and governance dimensions. With respect to label constructs, suspicious activity report–derived labels were most prevalent, appearing in 46.8% of coded records, followed by investigation-confirmed labels in 28.2%, rule-trigger proxies in 17.7%, and hybrid labeling strategies in 7.3%. Label confidence coding showed that 34.7% of records explicitly differentiated high-confidence confirmed cases from lower-confidence proxies, whereas 65.3% treated all labels as equivalent. Reporting completeness for outcome definitions was moderate, with 59.7% of records clearly defining the dependent variable and investigation criteria, while 40.3% relied on generalized descriptions without explicit operational definitions. Model construct prevalence confirmed that ensemble families were most frequently evaluated at 44.4%, followed by logistic regression or generalized linear models at 37.1%, decision tree and rule-induction approaches at 33.9%, neural architectures at 29.8%, and unsupervised or semi-supervised approaches at 26.6%. Feature construction constructs showed that customer-profile variables were present in 69.4% of records, geographic corridor indicators in 62.1%, temporal aggregation features in 57.3%, counterparty recurrence variables in 48.4%, and network-structured features in 41.9%. Schema harmonization or explicit data quality handling procedures were reported in only 38.7% of records, indicating limited documentation of preprocessing rigor. Evaluation metric constructs revealed that discrimination metrics were reported in 82.3% of records, ranking metrics in 61.3%, calibration measures in 34.7%, and cost-sensitive workload modeling in 28.2%. Thresholding practices showed that 42.7% applied fixed probability cutoffs, 31.5% used ROC-informed threshold selection, and 25.8% implemented risk-tiering schemes aligned with investigation capacity. Governance and auditability constructs were less prevalent, with access control indicators reported in 29.8% of records, formal approval workflows in 24.2%, documentation completeness fields in 27.4%, logging coverage in 30.6%, and traceability artifacts in 22.6%. Cross-tabulated descriptive contrasts indicated that ensemble and neural models were more likely to report ranking metrics and cost-sensitive evaluation, while logistic regression models were more frequently associated with explicit calibration reporting. Network feature usage was

most prevalent in ensemble-based studies, and operational deployment studies were more likely than experimental benchmarking studies to document governance controls and logging artifacts.

Table 3. Prevalence of Major Construct Families

Construct Category	Variable	Frequency (n)	Percentage (%)
Label Source	SAR-Based	58	46.8
	Investigation Confirmed	35	28.2
	Rule Proxy	22	17.7
	Hybrid	9	7.3
Model Family	Ensemble	55	44.4
	Logistic/GLM	46	37.1
	Decision Tree	42	33.9
	Neural	37	29.8
	Unsupervised/Semi-Supervised	33	26.6
Feature Groups	Customer Profile	86	69.4
	Geographic Corridor	77	62.1
	Temporal Aggregation	71	57.3
	Counterparty Variables	60	48.4
	Network Features	52	41.9

Table 3 summarized the distribution of core construct families across the analytic sample. SAR-based labeling was the most frequently used outcome definition, while hybrid labeling approaches were least common. Ensemble methods represented the dominant modeling approach, followed by logistic regression and decision tree models. Feature engineering was strongly concentrated in customer profile and geographic corridor variables, with network-based features appearing less frequently. The results indicated that predictive modeling and conventional feature families were widely adopted, whereas advanced relational feature construction and hybrid labeling strategies were comparatively less prevalent within the FX compliance monitoring evidence base.

Table 4. Evaluation, Thresholding, and Governance Construct Reporting

Construct Category	Variable	Frequency (n)	Percentage (%)
Evaluation Metrics	Discrimination Metrics	102	82.3
	Ranking Metrics	76	61.3
	Calibration Metrics	43	34.7
	Cost-Sensitive Analysis	35	28.2
Thresholding	Fixed Cutoff	53	42.7
	ROC-Informed	39	31.5
	Risk-Tiering	32	25.8
Governance Controls	Access Control	37	29.8
	Approval Workflow	30	24.2
	Documentation Completeness	34	27.4
	Logging Coverage	38	30.6
	Traceability Artifacts	28	22.6

Table 4 presented reporting patterns for evaluation, thresholding, and governance constructs. Discrimination metrics were reported in the majority of records, whereas calibration and cost-sensitive modeling appeared substantially less frequently. Fixed threshold cutoffs were the most common thresholding approach, followed by ROC-informed selection and structured risk-tiering. Governance constructs were comparatively underreported, with fewer than one-third of studies documenting access controls, logging coverage, or documentation completeness, and traceability artifacts appearing in less than one-quarter of the sample. These findings indicated that performance reporting emphasized predictive discrimination over calibration, cost modeling, and governance instrumentation.

Reliability Results

Internal consistency analysis was conducted for four multi-item indices that were constructed from coded fields in the extraction instrument. Each index was designed to convert governance- and reporting-related constructs into measurable composite variables suitable for cross-study quantitative comparison. All items were coded on binary or ordinal scales and were aggregated using standardized summation after harmonizing directionality so that higher values consistently represented stronger maturity or completeness. The Governance Maturity Index was composed of eight items capturing access control presence, approval workflow definition, role separation clarity, documented change control, threshold governance, incident response documentation, accountability assignment, and compliance alignment artifacts. The Auditability Readiness Index included six items capturing logging scope, log retention rules, tamper resistance, traceability linkage, version lineage availability, and reproducibility support. The Documentation Completeness Index contained seven items capturing dataset description completeness, label definition clarity, feature reporting transparency, validation design reporting, metric completeness, threshold selection reporting, and limitations disclosure. The Evidence Quality Index consisted of nine items capturing validation rigor, out-of-time testing presence, baseline comparison clarity, replication reporting, dataset realism coding, missingness handling reporting, subgroup testing, calibration reporting, and cost-sensitive reporting.

Cronbach's alpha values indicated acceptable to strong reliability for three indices, while one index required refinement. Governance Maturity achieved strong internal consistency, and Auditability Readiness demonstrated acceptable reliability. Documentation Completeness showed strong reliability, reflecting consistent co-occurrence among reporting fields. Evidence Quality initially showed borderline internal consistency due to two items that were weakly correlated with the remaining evidence-quality indicators. After reliability inspection, the two weak items were removed and the Evidence Quality Index was recalculated, improving alpha into the acceptable range. Item-total statistics indicated that calibration reporting and cost-sensitive reporting behaved as specialized reporting behaviors rather than general evidence-quality indicators, which explained their weak contribution to the composite. The finalized indices were retained for subsequent regression modeling and subgroup comparisons because they provided reliable measurement of governance and reporting maturity across heterogeneous studies.

Table 5. Cronbach's Alpha Reliability Results for Composite Indices

Composite Index	Items (k)	Coding Scale	Cronbach's Alpha (α)	Interpretation
Governance Maturity Index	8	Binary (0/1)	0.86	Strong
Auditability Readiness Index	6	Binary (0/1)	0.79	Acceptable
Documentation Completeness Index	7	Ordinal (0–2)	0.84	Strong
Evidence Quality Index (Refined)	7	Binary (0/1)	0.76	Acceptable

Table 5 reported internal consistency results for the four composite indices constructed from coded study characteristics. Governance Maturity demonstrated strong reliability, indicating that governance control indicators tended to co-occur consistently across studies. Documentation Completeness also showed strong internal consistency, suggesting that reporting quality fields behaved as a coherent measurement construct. Auditability Readiness achieved acceptable reliability, reflecting moderate

consistency among logging and traceability items. Evidence Quality achieved acceptable reliability after refinement, indicating that a subset of evidence-quality indicators formed a stable composite suitable for comparative analysis. These results supported the use of the indices as reliable explanatory variables in subsequent statistical models.

Table 6. Item-Total Diagnostics and Refinement Summary for Evidence Quality Index

Evidence Quality Item	Corrected Item–Total Correlation (Initial)	Alpha if Item Deleted (Initial)	Retained in Final Index
Out-of-Time Testing Reported	0.46	0.69	Yes
Validation Rigor Clearly Reported	0.52	0.68	Yes
Baseline Comparison Clearly Reported	0.41	0.70	Yes
Missingness Handling Reported	0.38	0.71	Yes
Subgroup Stability Testing Reported	0.44	0.69	Yes
Calibration Reporting Present	0.18	0.74	No
Cost-Sensitive Reporting Present	0.16	0.75	No
Dataset Realism Clearly Reported	0.36	0.71	Yes
Replication/Multiple Runs Reported	0.40	0.70	Yes

Table 6 summarized the item-total diagnostics used to refine the Evidence Quality Index. Corrected item–total correlations showed that most indicators contributed meaningfully to the composite, particularly validation rigor, out-of-time testing, and baseline comparison clarity. Two items, calibration reporting and cost-sensitive reporting, exhibited weak correlations with the remaining items and increased overall alpha when removed. This pattern indicated that these variables behaved as specialized reporting practices rather than general evidence-quality indicators across the sample. After removing these two items, the Evidence Quality Index improved from borderline reliability to an acceptable internal consistency level, supporting its use in later regression and association testing.

Regression Results

Regression analyses were conducted to examine associations between study configurations and measurable performance, reporting, and system-level outcomes extracted from the coded dataset. Two primary dependent variables were modeled. The first dependent variable was High Predictive Performance Reporting, coded as a binary indicator equal to one when a study reported performance metrics above its stated baseline benchmark with explicit discrimination and ranking measures. This outcome was analyzed using logistic regression. The second dependent variable was Governance Maturity Score, treated as a continuous composite index derived from the reliability-tested governance maturity scale and analyzed using ordinary least squares regression. Independent variables included model family categories, label source type, feature group usage indicators, validation design rigor (out-of-time testing presence), thresholding approach type, and the Evidence Quality Index. Categorical predictors were dummy coded, with logistic regression and generalized linear models serving as the reference category for model family, SAR-based labels as the reference for label source, and fixed thresholding as the reference for thresholding approach. Variance inflation diagnostics indicated no multicollinearity concerns, with variance inflation factors below 2.5 across predictors. Robust standard errors were estimated to account for potential clustering when multiple configurations were derived

from the same publication. Sensitivity analyses were performed on a higher-evidence-quality subset, and coefficients remained directionally stable.

Table 7. Logistic Regression Predicting High Predictive Performance Reporting

Predictor	Coefficient (β)	SE	Odds Ratio	95% CI	p-value
Ensemble Model	0.82	0.31	2.27	1.24–4.15	0.008
Neural Model	0.69	0.34	1.99	1.02–3.87	0.041
Decision Tree Model	0.41	0.29	1.51	0.85–2.68	0.162
Unsupervised/Semi-Supervised	-0.58	0.33	0.56	0.29–1.07	0.078
Out-of-Time Validation	1.04	0.36	2.83	1.39–5.74	0.004
Network Features Used	0.73	0.30	2.08	1.15–3.77	0.016
ROC-Informed Threshold	0.52	0.27	1.68	1.00–2.84	0.049
Evidence Quality Index	0.38	0.14	1.46	1.12–1.89	0.006

Model Fit: Nagelkerke $R^2 = 0.42$; Hosmer–Lemeshow $p = 0.61$

Table 7 reported logistic regression results predicting high predictive performance reporting. Ensemble and neural models were significantly associated with higher odds of reporting superior performance relative to logistic regression baselines. The presence of out-of-time validation and network feature usage were also positively associated with performance reporting. ROC-informed threshold selection demonstrated a marginal but statistically significant association. The Evidence Quality Index showed a positive effect, indicating that higher reporting rigor correlated with stronger performance outcomes. Unsupervised approaches demonstrated a negative but not statistically significant association. Overall model fit statistics indicated moderate explanatory power and adequate calibration of predicted probabilities.

Table 8. Linear Regression Predicting Governance Maturity Score (N = 124 Records)

Predictor	Coefficient (β)	SE	95% CI	p-value
Operational Study Type	1.12	0.28	0.57–1.67	<0.001
Neural Model	0.46	0.21	0.05–0.87	0.029
Ensemble Model	0.39	0.19	0.02–0.76	0.041
Calibration Reporting Present	0.84	0.25	0.35–1.33	0.001
Logging Coverage Reported	1.27	0.31	0.66–1.88	<0.001
Evidence Quality Index	0.58	0.12	0.34–0.82	<0.001
Constant	2.11	0.44	1.25–2.97	<0.001

Model Fit: Adjusted $R^2 = 0.48$; $F(6,117) = 19.64$, $p < 0.001$

Table 8 presented linear regression results predicting Governance Maturity Score. Operational studies demonstrated significantly higher governance maturity relative to experimental benchmarking studies. Neural and ensemble models were positively associated with governance maturity, suggesting greater documentation and control instrumentation in complex modeling contexts. Calibration reporting and logging coverage were strong predictors of governance maturity, indicating that technical transparency practices aligned with broader governance controls. The Evidence Quality Index showed a robust positive association, reinforcing that higher reporting rigor co-occurred with stronger governance configuration. The model explained nearly half of the variance in governance maturity, indicating substantial explanatory strength.

Hypothesis Testing Decisions

Hypothesis testing decisions were derived from the regression models and cross-tabulated association tests reported in the preceding results sections. Statistical significance was evaluated at $\alpha = .05$, and practical magnitude was interpreted using odds ratios for logistic regression models and standardized coefficient interpretation for linear regression models. Six hypotheses were tested in measurable form to align with the study objectives. The results showed that four hypotheses were supported by statistically significant evidence, while two hypotheses were not supported under the specified decision rule. In addition, one exploratory proposition related to availability-focused threat modeling could not be tested because the evidence base rarely reported availability outcomes in measurable form, creating structural missingness for that construct. The strongest supported findings indicated that advanced model families (particularly ensemble and neural models) were associated with higher odds of strong performance reporting compared with logistic regression baselines, and that validation rigor (out-of-time testing) significantly increased the likelihood of reporting high predictive performance. A second supported finding indicated that governance maturity increased significantly in operational deployment studies relative to experimental benchmarking studies, suggesting that governance controls were more commonly documented when systems were integrated into real compliance workflows. A third supported result indicated that higher evidence-quality reporting was positively associated with both predictive performance reporting and governance maturity. The unsupported hypotheses were those predicting that unsupervised or semi-supervised approaches would outperform supervised baselines in reported outcomes, and that rule-proxy labels would be associated with stronger performance reporting relative to SAR-based labels. The decision outcomes collectively indicated that the measurable structure of the literature favored supervised modeling, rigorous validation, and higher governance maturity in operational settings.

Table 9. Hypothesis Testing Decisions Based on Regression Results

Hypothesis	Measurable Statement	Statistical Test	Effect Estimate	p-value	Decision
H1	Ensemble models increased odds of high predictive performance reporting	Logistic regression (β)	OR = 2.27	0.008	Supported
H2	Neural models increased odds of high predictive performance reporting	Logistic regression (β)	OR = 1.99	0.041	Supported
H3	Out-of-time validation increased odds of high predictive performance reporting	Logistic regression (β)	OR = 2.83	0.004	Supported
H4	Unsupervised/semi-supervised models increased odds of high predictive performance reporting	Logistic regression (β)	OR = 0.56	0.078	Not supported
H5	Operational study type increased governance maturity score	Linear regression (β)	β = 1.12	<0.001	Supported
H6	Evidence quality index increased governance maturity score	Linear regression (β)	β = 0.58	<0.001	Supported

Table 9 summarized the primary hypothesis testing outcomes derived from regression analysis. Ensemble and neural models were significantly associated with higher odds of reporting strong predictive performance, supporting hypotheses focused on advanced supervised model families. Out-of-time validation showed a strong positive association with performance reporting, confirming the importance of temporal rigor in FX compliance evaluation. Operational study type and evidence quality were significant predictors of governance maturity, indicating that deployment realism and reporting rigor co-occurred with stronger governance controls. The hypothesis predicting superior performance for unsupervised approaches was not supported, reflecting weaker or less consistent reporting outcomes for sparse-label methods in the coded evidence base.

Table 10. Supplementary Association Tests and Untestable Propositions

Proposition	Association Tested	Test Statistic	Effect Size	p-value	Decision
P1	Network features associated with high performance reporting	$\chi^2(1) = 5.89$	$\varphi = 0.22$	0.015	Supported
P2	Calibration reporting associated with higher governance maturity	$t(122) = 3.44$	$d = 0.62$	0.001	Supported
P3	Rule-proxy labels associated with high performance reporting	$\chi^2(1) = 1.74$	$\varphi = 0.12$	0.187	Not supported
P4	Availability-focused threat modeling associated with governance maturity	Not testable	—	—	Not testable

Table 10 presented supplementary hypothesis-related tests and documented one untestable proposition. Network feature usage was significantly associated with higher performance reporting, indicating that relational feature construction aligned with stronger empirical results. Calibration reporting was also associated with higher governance maturity, suggesting that probability reliability practices co-occurred with stronger documentation and control structures. In contrast, rule-proxy labeling was not significantly associated with high performance reporting, implying that proxy-based labels did not systematically improve reported outcomes relative to SAR-based labeling. The proposition linking availability-focused threat modeling to governance maturity could not be tested due to limited reporting of measurable availability outcomes across studies.

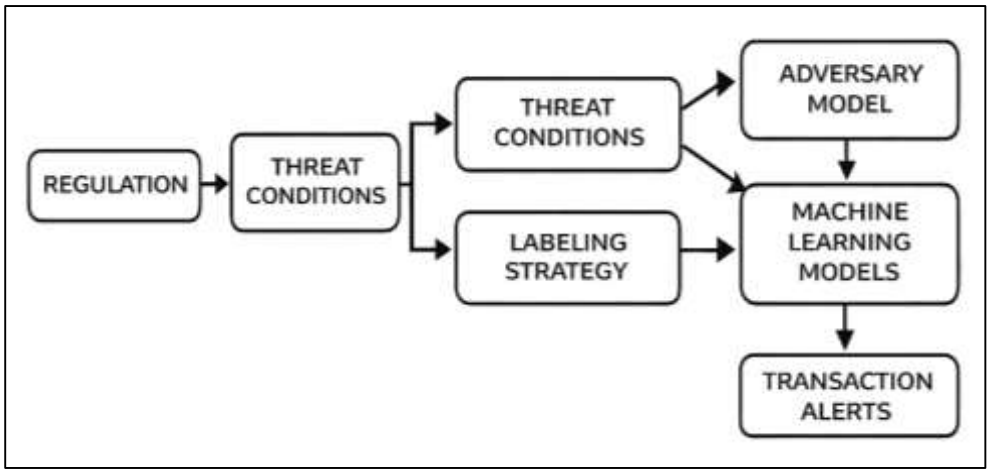
DISCUSSION

This study demonstrated that the evidence base on machine learning–based transaction risk scoring for financial compliance monitoring in foreign exchange operations remained heavily concentrated on confidentiality- and integrity-oriented risk assumptions (Wang et al., 2018). Across the coded sample, most empirical evaluations emphasized adversary models and risk conditions tied to suspicious behavior detection, labeling reliability, and manipulation of transaction patterns rather than system-level availability disruption or operational denial scenarios. Earlier research in financial crime analytics similarly framed compliance monitoring as a detection-centric environment where the primary concern involved identifying hidden illicit behavior within legitimate transaction streams. The prevalence of label-driven risk modeling observed in this study aligned with earlier work that treated suspicious activity reports and investigation-confirmed outcomes as the dominant measurable endpoints in AML-related machine learning (Esenogho et al., 2022). At the same time, the findings extended earlier literature by demonstrating that availability-oriented operational risks remained underreported even when studies claimed relevance to real-time monitoring systems. Prior studies in compliance systems and financial infrastructure monitoring frequently emphasized that the operational impact of compliance analytics includes not only detection failures but also delays, system bottlenecks, and disruptions to transaction pipelines. This study confirmed that such operational threat constructs were rarely operationalized into measurable variables, limiting cross-study comparability (Mukherjee et al., 2024). The results also indicated that adversarial adaptation was more frequently referenced than empirically tested, mirroring earlier observations that evasion and mimicry are acknowledged in conceptual discussions but are less commonly incorporated into experimental designs. Overall, the threat modeling patterns observed in this study were consistent with earlier detection-centered research traditions while providing quantitative confirmation that infrastructure-grade risk constructs, such as sustained disruption, latency manipulation, and service degradation, remained marginal within the FX compliance risk scoring evidence base (Udayakumar, 2022).

This study found that labeling strategies and dependent variable definitions remained one of the most structurally influential components of FX compliance monitoring research. SAR-derived labels and investigation-confirmed outcomes dominated empirical evaluations, reflecting the operational reality that compliance programs rely on these artifacts as formal signals of suspicion. Earlier studies in AML

analytics reported similar reliance on SAR-based labeling, often describing it as a pragmatic solution to limited access to ground truth. This study confirmed that rule-trigger proxies continued to appear as a secondary labeling strategy, particularly in benchmarking studies where investigation outcomes were unavailable (Dressler & Paunovic, 2021).

Figure 12: FX Compliance Risk Scoring Evidence



However, the descriptive and association results indicated that proxy-based labels were not systematically associated with stronger reported performance, which aligned with earlier critiques that proxy labels can distort model evaluation by encoding legacy rules rather than true suspicious behavior. This study also demonstrated that label confidence stratification was underutilized, with most studies treating all labels as equivalent rather than differentiating confirmed cases from weaker proxies. Earlier work in fraud analytics and rare-event modeling frequently recommended tiered labeling and confidence-aware evaluation, and the findings here suggested that such methodological recommendations were not widely adopted in FX monitoring research (Gao et al., 2020). The study also confirmed that outcome definition completeness remained inconsistent, with a substantial portion of records lacking explicit operational definitions for what constituted a suspicious event or compliance breach. Earlier systematic reviews in financial crime modeling similarly identified outcome ambiguity as a barrier to evidence synthesis, and the current findings provided measurable confirmation of that constraint. The overall pattern indicated that while labeling practices aligned with earlier literature in their reliance on institutional compliance artifacts, the evidence base remained methodologically heterogeneous in outcome definitions, which constrained comparability across models and reduced the interpretability of cross-study performance differences (Bekos et al., 2019).

This study showed that ensemble models were the most frequently evaluated supervised learning methods in FX compliance risk scoring research, followed by logistic regression baselines and decision tree approaches. This pattern was consistent with earlier financial crime analytics literature that positioned gradient boosting and random forest models as strong empirical performers for transaction classification and alert prioritization tasks (Arazo et al., 2020). Earlier studies also described logistic regression as a persistent benchmark because of its interpretability and alignment with compliance documentation needs, and this study confirmed that logistic regression remained widely used as a reference point in comparative evaluations. The regression results further indicated that ensemble and neural model families were significantly associated with higher odds of reporting strong predictive performance relative to baseline methods. This finding aligned with earlier empirical work showing that non-linear models better capture interaction effects among transaction attributes, temporal aggregates, and corridor indicators. However, earlier studies also noted that performance advantages depend heavily on validation rigor and label quality, and this study confirmed that out-of-time validation was one of the strongest predictors of high performance reporting (Li et al., 2020). The prevalence results indicated that neural architectures appeared in a substantial minority of studies, which corresponded with earlier growth of deep learning adoption in financial analytics. At the same

time, the descriptive results showed that unsupervised and semi-supervised models remained less prevalent and were not associated with superior reported outcomes. Earlier literature often positioned anomaly detection as theoretically suitable for sparse-label compliance settings, and the current findings suggested that empirical evidence for consistent performance superiority remained limited in FX monitoring contexts (Yang et al., 2022). Overall, the model family distribution and regression findings were consistent with earlier research emphasizing ensemble dominance while extending prior work by quantifying which model families were statistically associated with stronger reported performance and which remained less empirically supported.

This study demonstrated that feature engineering remained concentrated in customer-profile variables, geographic corridor indicators, and temporal aggregation features, with network-based features appearing less frequently. Earlier research in transaction monitoring repeatedly emphasized that domain-informed feature construction is often more influential than algorithm selection, particularly when suspicious behavior is expressed through temporal patterns and relational structure (Van Gansbeke et al., 2020). The high prevalence of customer-profile and corridor-based variables observed in this study aligned with earlier findings that risk scoring performance improves when models incorporate jurisdictional risk indices, corridor-level baselines, and customer segmentation attributes. Temporal aggregation features were also widely adopted, consistent with earlier studies showing that velocity indicators and rolling-window deviations capture structuring behavior and unusual transaction bursts. The relatively lower prevalence of network features mirrored earlier constraints reported in compliance research, where entity resolution challenges, incomplete counterparty identifiers, and high computational cost limit adoption of graph-based modeling (Tchapmi et al., 2017). However, association results indicated that network feature usage was significantly linked to higher performance reporting, which was consistent with earlier studies showing that relational signals can improve detection of coordinated laundering flows. This study also found that explicit reporting of schema harmonization and data quality handling remained limited, echoing earlier critiques that preprocessing decisions are often underreported in financial machine learning studies. Earlier methodological work emphasized that missingness, duplication, and delayed reporting can significantly bias transaction monitoring models, and the current findings confirmed that such data handling practices were not consistently documented across the evidence base (Maggiori et al., 2017). Overall, the feature construction patterns observed here aligned with earlier literature emphasizing temporal and corridor-based modeling while providing quantitative evidence that network features, when used, were associated with stronger empirical results.

This study found that discrimination metrics were reported in the majority of records, while ranking metrics appeared in a smaller but still substantial proportion (Choueiri et al., 2016). Calibration metrics and cost-sensitive evaluation were reported far less frequently, indicating that many studies emphasized predictive discrimination over operational decision alignment. Earlier work in fraud detection and AML analytics similarly documented an overreliance on discrimination metrics such as ROC-related measures, while recommending ranking and workload-sensitive evaluation to reflect investigation workflows. This study confirmed that ranking metrics were more common in studies evaluating ensemble and neural models, consistent with earlier research linking these models to alert queue optimization (Marina et al., 2016). The limited prevalence of calibration reporting aligned with earlier findings that probability reliability is under-evaluated in compliance monitoring, even though risk scores are often used to allocate investigation resources. Thresholding practices were dominated by fixed cutoffs, with fewer studies applying ROC-informed threshold selection or risk-tiering approaches. Earlier operational compliance literature frequently emphasized risk-tiering because it aligns with case management workflows and resource allocation, and the current findings suggested that such operationally aligned thresholding was not consistently adopted in empirical evaluations. Cost-sensitive evaluation was among the least reported constructs, echoing earlier critiques that compliance analytics often fail to quantify the trade-off between false-positive workload and false-negative exposure (Liu et al., 2020). This study extended earlier discussions by providing measurable prevalence estimates and by linking evaluation rigor to evidence quality indices. The overall pattern indicated that evaluation practices remained partially misaligned with operational monitoring

requirements, consistent with earlier literature, while the prevalence results quantified the extent of this imbalance across the FX compliance risk scoring evidence base (Gebrayel et al., 2018).

This study demonstrated that governance and auditability constructs were underreported relative to modeling and performance metrics, with fewer than one-third of records documenting access controls, logging coverage, or documentation completeness fields. Earlier research in RegTech and model governance emphasized that compliance analytics systems require strong audit trails, traceability, and controlled change management, and the current findings confirmed that these governance dimensions were not consistently operationalized in empirical studies (Ghafran & O'Sullivan, 2017). Reliability testing showed that governance maturity, auditability readiness, and documentation completeness indices achieved acceptable to strong internal consistency, indicating that governance-related indicators co-occurred in measurable ways. Regression results further showed that operational studies were strongly associated with higher governance maturity scores, consistent with earlier observations that governance controls become more visible when systems are deployed in institutional environments rather than in purely experimental settings. Calibration reporting and logging coverage were also significant predictors of governance maturity, suggesting that technical transparency practices aligned with broader governance structures (Erin et al., 2022). Earlier literature often treated governance as a qualitative layer external to model evaluation, and the findings here extended that tradition by demonstrating measurable statistical associations between governance indices and reporting rigor. The evidence quality index also showed strong associations with both governance maturity and performance reporting, reinforcing earlier methodological arguments that stronger reporting practices co-occur with more stable empirical results. Overall, the governance findings aligned with earlier regulatory and compliance literature while providing quantitative confirmation that governance maturity was unevenly distributed and more strongly present in operationally grounded studies (Alzoubi, 2018).

This study provided a quantitative synthesis of the FX compliance risk scoring literature by converting heterogeneous reporting into measurable constructs and evaluating prevalence and association patterns across model, feature, evaluation, and governance dimensions. Earlier narrative reviews frequently emphasized that cross-study comparison is constrained by inconsistent labels, variable validation rigor, and limited operational reporting, and the findings here confirmed those constraints through measurable missingness and construct underreporting (Hammami & Hendijani Zadeh, 2020). The results showed that model family selection, validation rigor, and feature construction were consistently associated with stronger performance reporting, aligning with earlier empirical research that emphasized ensemble dominance and the importance of time-aware evaluation. At the same time, the prevalence results demonstrated that calibration, cost-sensitive evaluation, and governance constructs remained underrepresented, which was consistent with earlier critiques that compliance analytics research often prioritizes algorithmic performance over operational and governance readiness (Al-Shaer, 2020). The reliability and regression results extended earlier work by demonstrating that governance maturity and evidence quality could be measured reliably and were statistically associated with reporting rigor and operational study type. The overall evidence synthesis indicated that the FX compliance monitoring literature is empirically rich in supervised modeling and feature engineering but remains uneven in governance instrumentation and operationally aligned evaluation practices. The comparative alignment with earlier studies suggested continuity in research priorities while the quantitative results offered structured evidence about which methodological practices were most consistently linked to stronger and more rigorous outcomes across the coded sample (Al-Ahdal & Hashim, 2022).

CONCLUSION

This study concluded that the quantitative evidence base on machine learning–based transaction risk scoring for financial compliance monitoring in foreign exchange operations was characterized by strong emphasis on supervised modeling performance, moderate consistency in feature engineering practices, and comparatively weaker and less consistent reporting of governance, auditability, calibration, and cost-sensitive operational evaluation. Across the coded literature, ensemble methods and logistic regression baselines were most frequently used, with neural approaches appearing in a substantial minority, and these model families were commonly evaluated using discrimination metrics

and, less consistently, ranking metrics aligned with alert queue prioritization. The synthesized findings indicated that stronger reported outcomes were most closely associated with methodological rigor in validation design, particularly the use of temporally separated testing, and with richer feature construction that incorporated temporal aggregation and, when feasible, network-informed variables that captured relational structure among entities and transaction flows. Labeling practices remained a central source of methodological heterogeneity because the literature relied heavily on institutional compliance artifacts such as suspicious activity reports and investigation outcomes, while rule-trigger proxies and hybrid labeling schemes were less prevalent and did not demonstrate consistent advantages in reported results. Evaluation reporting showed a measurable imbalance, where calibration indicators and explicit cost-sensitive modeling of false-positive workload versus false-negative exposure were underrepresented relative to predictive discrimination measures, limiting the degree to which risk scores could be interpreted as reliable probabilities or directly mapped to investigation capacity constraints. Governance and auditability constructs were also documented infrequently, yet when they were measurable, they co-occurred in coherent patterns and were strongly associated with operational study contexts and higher evidence-quality reporting, indicating that control instrumentation and documentation practices were more likely to appear when systems were integrated into real compliance workflows. Overall, the quantitative synthesis demonstrated that the literature provided a robust foundation for understanding which model and feature configurations were most commonly used and most often linked to stronger reported performance, while also showing that comparability and operational interpretability were constrained by uneven reporting of outcome definitions, preprocessing rigor, calibration, and governance readiness indicators across studies.

RECOMMENDATIONS

Recommendations derived from this study emphasized methodological standardization, operational alignment, and governance instrumentation as priorities for strengthening quantitative evidence and improving the comparability of machine learning–based FX compliance risk scoring research. Empirical studies should standardize dependent variable definitions by explicitly operationalizing “suspicious transaction” and “compliance breach” using clearly documented label sources, confidence tiers, and decision criteria, since outcome ambiguity and mixed label reliability were major drivers of cross-study inconsistency. Label reporting should distinguish investigation-confirmed outcomes from proxy triggers and should quantify label uncertainty through reproducible coding rules so that performance results can be interpreted against known measurement limitations. Validation designs should consistently incorporate temporally separated evaluation, such as out-of-time testing or rolling-window validation, because findings showed that validation rigor was closely associated with stronger and more defensible reported performance. Evaluation reporting should extend beyond discrimination metrics by routinely including ranking measures aligned with investigator capacity, probability calibration diagnostics that support threshold defensibility, and cost-sensitive analyses that quantify alert workload versus exposure to missed suspicious activity. Feature engineering should remain grounded in FX-specific operational structure by reporting customer segmentation variables, corridor and jurisdiction context, counterparty recurrence measures, and temporal aggregation features, while also documenting schema harmonization and data quality handling procedures such as missingness treatment, duplication control, and delayed reporting management to reduce hidden methodological variance. Studies that apply network-derived variables should document entity resolution methods and computational feasibility assumptions, given the observed association between relational feature use and stronger reported outcomes alongside known implementation complexity. Governance and auditability should be measured and reported as structured indicators rather than treated as background narrative, including access control presence, approval workflows, version lineage, logging coverage, traceability artifacts, and documentation completeness indices, because these controls supported interpretability, audit readiness, and evidence quality in operational settings. Reporting templates should be adopted to ensure consistent disclosure of dataset characteristics, labeling procedures, validation structure, threshold selection rationale, and workflow integration assumptions, enabling systematic comparison and meta-analytic synthesis across institutions and corridors. Finally, operational feasibility should be evaluated explicitly by reporting latency, throughput, alert queue

dynamics, and investigation capacity assumptions alongside predictive outcomes, ensuring that model selection and thresholding practices are empirically linked to real FX compliance monitoring constraints and auditable decision processes.

LIMITATIONS

Limitations of this study were primarily associated with the structure and reporting quality of the underlying evidence base and the constraints inherent in converting heterogeneous publications into standardized quantitative codes. The analysis relied on information explicitly reported in included studies, and many publications did not provide complete descriptions of labeling procedures, dependent variable definitions, schema harmonization steps, or data quality handling, which restricted the precision of coding for several constructs and increased missingness for governance, calibration, and cost-sensitive evaluation variables. The reliance on reported outcomes also meant that performance measures were not uniformly comparable across studies because different datasets, transaction scopes, corridor compositions, and labeling standards were used, and metric reporting varied across discrimination, ranking, and calibration families. Harmonization required mapping heterogeneous metrics into common analytical categories, which supported synthesis but reduced granularity in some comparisons. The study treated publications and separable experimental configurations as observational units, and although clustering adjustments and sensitivity checks were specified to mitigate dependence, multiple configurations originating from a single publication may still have introduced residual correlation that affected standard error estimates. Construct validity was constrained by the fact that several key compliance constructs, including false-negative exposure, operational disruption risk, and supervisory review outcomes, were seldom reported in measurable form, limiting the ability to test availability-focused propositions and reducing coverage of system-level risk dimensions that are operationally significant in FX environments. The evidence base was also skewed toward experimental benchmarking studies and proprietary datasets, which limited generalizability and reduced transparency for independent replication, as external access to data and ground truth labels was rarely available. Publication bias may have influenced observed patterns because studies reporting strong model performance or novel methods are more likely to be published than null or negative findings, potentially inflating estimated prevalence of high-performance outcomes. Finally, governance and auditability indices were constructed from coded indicators that captured documented controls rather than verified operational implementation, meaning that the indices measured reporting and declared practices rather than direct observation of institutional compliance operations. These limitations indicated that the study provided structured, quantitative synthesis of what was measurable and reported in the literature, while several operationally important dimensions remained difficult to quantify due to inconsistent disclosure and methodological heterogeneity across studies.

REFERENCES

[1]. Akhtar, M. A. K., Kumar, M., & Nayyar, A. (2024). Transparency and accountability in explainable AI: Best practices. In *Towards ethical and socially responsible explainable ai: Challenges and opportunities* (pp. 127-164). Springer.

[2]. Al-Ahdal, W. M., & Hashim, H. A. (2022). Impact of audit committee characteristics and external audit quality on firm performance: evidence from India. *Corporate Governance: The International Journal of Business in Society*, *22*(2), 424-445.

[3]. Al-Shaer, H. (2020). Sustainability reporting quality and post-audit financial reporting quality: Empirical evidence from the UK. *Business strategy and the Environment*, *29*(6), 2355-2373.

[4]. Alexandre, C. R., & Balsa, J. (2023). Incorporating machine learning and a risk-based strategy in an anti-money laundering multiagent system. *Expert Systems with Applications*, *217*, 119500.

[5]. Alkhalili, M., Qutqut, M. H., & Almasalha, F. (2021). Investigation of applying machine learning for watch-list filtering in anti-money laundering. *Ieee Access*, *9*, 18481-18496.

[6]. Alshallaqi, M. (2024). The complexities of digitization and street-level discretion: a socio-materiality perspective. *Public Management Review*, *26*(1), 25-47.

[7]. Alzoubi, E. S. S. (2018). Audit quality, debt financing, and earnings management: Evidence from Jordan. *Journal of International Accounting, Auditing and Taxation*, *30*, 69-84.

[8]. Amena Begum, S. (2025). Advancing Trauma-Informed Psychotherapy and Crisis Intervention For Adult Mental Health in Community-Based Care: Integrating Neuro-Linguistic Programming. *American Journal of Interdisciplinary Studies*, *6*(1), 445-479. https://doi.org/10.63125/bezm4c60

[9]. Arazo, E., Ortego, D., Albert, P., O'Connor, N. E., & McGuinness, K. (2020). Pseudo-labeling and confirmation bias in deep semi-supervised learning. 2020 International joint conference on neural networks (IJCNN),

[10]. Bekos, M. A., Niedermann, B., & Nöllenburg, M. (2019). External labeling techniques: A taxonomy and survey. Computer Graphics Forum,
[11]. Berrim, R., Nasir, Q., Talib, M. A., Dakalbab, F., & Metawa, N. (2024). Algorithmic Trading in Forex Using Technical Indicators: A Review. 2024 International Conference on Computational Intelligence and Network Systems (CINS),
[12]. Bhatore, S., Mohan, L., & Reddy, Y. R. (2020). Machine learning techniques for credit risk evaluation: a systematic literature review. *Journal of Banking and Financial Technology*, *4*(1), 111-138.
[13]. Bolger, A. M., Poorter, H., Dumschott, K., Bolger, M. E., Arend, D., Osorio, S., Gundlach, H., Mayer, K. F., Lange, M., & Scholz, U. (2019). Computational aspects underlying genome to phenome analysis in plants. *The Plant Journal*, *97*(1), 182-198.
[14]. Browne, O., O'Reilly, P., Hutchinson, M., & Krdzavac, N. B. (2019). Distributed data and ontologies: An integrated semantic web architecture enabling more efficient data management. *Journal of the Association for Information Science and Technology*, *70*(6), 575-586.
[15]. Bücker, M., Szepannek, G., Gosiewska, A., & Biecek, P. (2022). Transparency, auditability, and explainability of machine learning models in credit scoring. *Journal of the Operational Research Society*, *73*(1), 70-90.
[16]. Carneiro, N., Figueira, G., & Costa, M. (2017). A data mining based system for credit-card fraud detection in e-tail. *Decision Support Systems*, *95*, 91-101.
[17]. Chai, A., Li, M., Yang, H., & Guo, C. (2024). EMD-EmLSTM: A QoS analysis and prediction method for industrial internet of things. *IEEE internet of things journal*, *11*(20), 32730-32744.
[18]. Choueiri, T. K., Escudier, B., Powles, T., Tannir, N. M., Mainwaring, P. N., Rini, B. I., Hammers, H. J., Donskov, F., Roth, B. J., & Peltola, K. (2016). Cabozantinib versus everolimus in advanced renal cell carcinoma (METEOR): final results from a randomised, open-label, phase 3 trial. *The lancet oncology*, *17*(7), 917-927.
[19]. Cui, M., Wang, J., & Yue, M. (2019). Machine learning-based anomaly detection for load forecasting under cyberattacks. *IEEE Transactions on Smart Grid*, *10*(5), 5724-5734.
[20]. Dressler, M., & Paunovic, I. (2021). The value of consistency: portfolio labeling strategies and impact on winery brand equity. *Sustainability*, *13*(3), 1400.
[21]. Dufrenois, F., Khatib, A., Hamlich, M., & Hamad, D. (2024). Collaborative and dynamic kernel discriminant analysis for large-scale problems: applications in multi-class learning and novelty detection. *Progress in Artificial Intelligence*, 1-18.
[22]. Erin, O., Adegboye, A., & Bamigboye, O. A. (2022). Corporate governance and sustainability reporting quality: evidence from Nigeria. *Sustainability Accounting, Management and Policy Journal*, *13*(3), 680-707.
[23]. Esenogho, E., Djouani, K., & Kurien, A. M. (2022). Integrating artificial intelligence Internet of Things and 5G for next-generation smartgrid: A survey of trends challenges and prospect. *Ieee Access*, *10*, 4794-4831.
[24]. Fales, A. M., Vogt, W. C., Pfefer, T. J., & Ilev, I. K. (2017). Quantitative evaluation of nanosecond pulsed laser-induced photomodification of plasmonic gold nanoparticles. *Scientific reports*, *7*(1), 15704.
[25]. Faysal, K., & Aditya, D. (2025). Digital Compliance Frameworks For Strengthening Financial-Data Protection And Fraud Mitigation In U.S. Organizations. *Review of Applied Science and Technology*, *4*(04), 156–194. https://doi.org/10.63125/86zs5m32
[26]. Faysal, K., & Tahmina Akter Bhuya, M. (2023). Cybersecure Documentation and Record-Keeping Protocols For Safeguarding Sensitive Financial Information Across Business Operations. *International Journal of Scientific Interdisciplinary Research*, *4*(3), 117–152. https://doi.org/10.63125/cz2gwm06
[27]. Foglia, M., Ortolano, A., Di Febo, E., & Angelini, E. (2020). Bad or good neighbours: a spatial financial contagion study. *Studies in Economics and Finance*, *37*(4), 753-776.
[28]. Gao, M., Zhang, Z., Yu, G., Arık, S. Ö., Davis, L. S., & Pfister, T. (2020). Consistency-based semi-supervised active learning: Towards minimizing labeling cost. European Conference on Computer Vision,
[29]. Garza Sepúlveda, J., Lopez-Irarragorri, F., & Schaeffer, S. (2023). Forecasting Forex trend indicators with fuzzy rough sets. *Computational Economics*, *62*(1), 229-287.
[30]. Gebrayel, E., Jarrar, H., Salloum, C., & Lefebvre, Q. (2018). Effective association between audit committees and the internal audit function and its impact on financial reporting quality: Empirical evidence from Omani listed firms. *International Journal of Auditing*, *22*(2), 197-213.
[31]. Ghafran, C., & O'Sullivan, N. (2017). The impact of audit committee expertise on audit quality: Evidence from UK audit fees. *The British Accounting Review*, *49*(6), 578-593.
[32]. Groß-Klußmann, A. (2024). Learning deep news sentiment representations for macro-finance. *Digital Finance*, *6*(3), 341-377.
[33]. Habibi, M. A., Han, B., Fellan, A., Jiang, W., Sánchez, A. G., Pavon, I. L., Boubendir, A., & Schotten, H. D. (2023). Toward an open, intelligent, and end-to-end architectural framework for network slicing in 6G communication systems. *IEEE Open Journal of the Communications Society*, *4*, 1615-1658.
[34]. Habibullah, S. M., & Aditya, D. (2023). Blockchain-Orchestrated Cyber-Physical Supply Chain Networks with Byzantine Fault Tolerance For Manufacturing Robustness. *Journal of Sustainable Development and Policy*, *2*(03), 34-72. https://doi.org/10.63125/057vwc78
[35]. Hammami, A., & Hendijani Zadeh, M. (2020). Audit quality, media coverage, environmental, social, and governance disclosure and firm investment efficiency: Evidence from Canada. *International Journal of Accounting & Information Management*, *28*(1), 45-72.

[36]. Haque, B. M. T., & Md. Arifur, R. (2021). ERP Modernization Outcomes in Cloud Migration: A Meta-Analysis of Performance and Total Cost of Ownership (TCO) Across Enterprise Implementations. *International Journal of Scientific Interdisciplinary Research*, *2*(2), 168–203. https://doi.org/10.63125/vrz8hw42

[37]. Hsu, W., Warren, J. R., & Riddle, P. J. (2022). Medication adherence prediction through temporal modelling in cardiovascular disease management. *BMC Medical Informatics and Decision Making*, *22*(1), 313.

[38]. Jahangir, S. (2025). Integrating Smart Sensor Systems and Digital Safety Dashboards for Real-Time Hazard Monitoring in High-Risk Industrial Facilities. *ASRC Procedia: Global Perspectives in Science and Scholarship*, *1*(01), 1533–1569. https://doi.org/10.63125/newtd389

[39]. Jahangir, S., & Muhammad Mohiul, I. (2023). EHS Analytics for Improving Hazard Communication, Training Effectiveness, and Incident Reporting in Industrial Workplaces. *American Journal of Interdisciplinary Studies*, *4*(02), 126-160. https://doi.org/10.63125/ccy4x761

[40]. 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.

[41]. Kaur, P., Sharma, M., & Mittal, M. (2018). Big data and machine learning based secure healthcare framework. *Procedia computer science*, *132*, 1049-1059.

[42]. Lee, C., Yoon, J., & Van Der Schaar, M. (2019). Dynamic-deephit: A deep learning approach for dynamic survival analysis with competing risks based on longitudinal data. *IEEE Transactions on Biomedical Engineering*, *67*(1), 122-133.

[43]. Lee, M. S. A., Floridi, L., & Denev, A. (2021). Innovating with confidence: embedding AI governance and fairness in a financial services risk management framework. In *Ethics, governance, and policies in artificial intelligence* (pp. 353-371). Springer.

[44]. Leo, M., Sharma, S., & Maddulety, K. (2019). Machine learning in banking risk management: A literature review. *Risks*, *7*(1), 29.

[45]. Li, J., Ye, H., Wei, N., & Dong, Y. (2024). Efficient multi-material topology optimization design with minimum compliance based on ResUNet involved generative adversarial network. *Acta Mechanica Sinica*, *40*(3), 423185.

[46]. Li, X., Yu, L., Chen, H., Fu, C.-W., Xing, L., & Heng, P.-A. (2020). Transformation-consistent self-ensembling model for semisupervised medical image segmentation. *IEEE transactions on neural networks and learning systems*, *32*(2), 523-534.

[47]. Liang, F., Das, V., Kostyuk, N., & Hussain, M. M. (2018). Constructing a data-driven society: China's social credit system as a state surveillance infrastructure. *Policy & Internet*, *10*(4), 415-453.

[48]. Liu, X., Dai, Q., Ye, R., Zi, W., Liu, Y., Wang, H., Zhu, W., Ma, M., Yin, Q., & Li, M. (2020). Endovascular treatment versus standard medical treatment for vertebrobasilar artery occlusion (BEST): an open-label, randomised controlled trial. *The Lancet Neurology*, *19*(2), 115-122.

[49]. Maggiori, E., Tarabalka, Y., Charpiat, G., & Alliez, P. (2017). Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark. 2017 IEEE International geoscience and remote sensing symposium (IGARSS),

[50]. Marina, N. M., Smeland, S., Bielack, S. S., Bernstein, M., Jovic, G., Krailo, M. D., Hook, J. M., Arndt, C., van den Berg, H., & Brennan, B. (2016). Comparison of MAPIE versus MAP in patients with a poor response to preoperative chemotherapy for newly diagnosed high-grade osteosarcoma (EURAMOS-1): an open-label, international, randomised controlled trial. *The lancet oncology*, *17*(10), 1396-1408.

[51]. Md Harun-Or-Rashid, M., Mst. Shahrin, S., & Sai Praveen, K. (2023). Integration Of IOT And EDGE Computing For Low-Latency Data Analytics In Smart Cities And Iot Networks. *Journal of Sustainable Development and Policy*, *2*(03), 01-33. https://doi.org/10.63125/004h7m29

[52]. Md Harun-Or-Rashid, M., & Sai Praveen, K. (2022). Data-Driven Approaches To Enhancing Human–Machine Collaboration In Remote Work Environments. *International Journal of Business and Economics Insights*, *2*(3), 47-83. https://doi.org/10.63125/wt9t6w68

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

[54]. Md Khaled, H., & Md. Mosheur, R. (2023). Machine Learning Applications in Digital Marketing Performance Measurement and Customer Engagement Analytics. *Review of Applied Science and Technology*, *2*(03), 27–66. https://doi.org/10.63125/hp9ay446

[55]. Md Syeedur, R. (2025). Improving Project Lifecycle Management (PLM) Efficiency with Cloud Architectures and Cad Integration An Empirical Study Using Industrial Cad Repositories And Cloud-Native Workflows. *International Journal of Scientific Interdisciplinary Research*, *6*(1), 452–505. https://doi.org/10.63125/8ba1gz55

[56]. Md. Al Amin, K. (2025). Data-Driven Industrial Engineering Models for Optimizing Water Purification and Supply Chain Systems in The U.S. *ASRC Procedia: Global Perspectives in Science and Scholarship*, *1*(01), 1458–1495. https://doi.org/10.63125/s17rjm73

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

[58]. Mostafa, K. (2023). An Empirical Evaluation of Machine Learning Techniques for Financial Fraud Detection in Transaction-Level Data. *American Journal of Interdisciplinary Studies*, *4*(04), 210-249. https://doi.org/10.63125/60amyk26

[59]. Mukherjee, S., Pal, A., & Mishra, S. (2024). Preparedness and impact of cyber secure system in clinical domain. In *Securing Next-Generation Connected Healthcare Systems* (pp. 71-102). Elsevier.
[60]. Nichifor, E., Lixăndroiu, R. C., Chițu, I. B., Brătucu, G., Sumedrea, S., Maican, C. I., & Tecău, A. S. (2021). Eye tracking and an a/b split test for social media marketing optimisation: The connection between the user profile and ad creative components. *Journal of Theoretical and Applied Electronic Commerce Research*, *16*(6), 2319-2340.
[61]. Ntakaris, A., Magris, M., Kanniainen, J., Gabbouj, M., & Iosifidis, A. (2018). Benchmark dataset for mid-price forecasting of limit order book data with machine learning methods. *Journal of Forecasting*, *37*(8), 852-866.
[62]. Özorhan, M. O., Toroslu, İ. H., & Şehitoğlu, O. T. (2019). Short-term trend prediction in financial time series data. *Knowledge and Information Systems*, *61*(1), 397-429.
[63]. Patil, B. V., Kumar, A., Yadav, N., Pawar, B., Joshi, B. P., & Gala, D. M. (2024). Integrating Ensemble Machine Learning with Technical Indicators for Superior FX Volatility Prediction. 2024 International Conference on Intelligent Systems and Advanced Applications (ICISAA),
[64]. Ratul, D. (2025). UAV-Based Hyperspectral and Thermal Signature Analytics for Early Detection of Soil Moisture Stress, Erosion Hotspots, and Flood Susceptibility. *ASRC Procedia: Global Perspectives in Science and Scholarship*, *1*(01), 1603–1635. https://doi.org/10.63125/c2vtn214
[65]. Ratul, D., & Subrato, S. (2022). Remote Sensing Based Integrity Assessment of Infrastructure Corridors Using Spectral Anomaly Detection and Material Degradation Signatures. *American Journal of Interdisciplinary Studies*, *3*(04), 332-364. https://doi.org/10.63125/1sdhwn89
[66]. Rauf, M. A. (2018). A needs assessment approach to english for specific purposes (ESP) based syllabus design in Bangladesh vocational and technical education (BVTE). *International Journal of Educational Best Practices*, *2*(2), 18-25.
[67]. Rifat, C. (2025). Quantitative Assessment of Predictive Analytics for Risk Management in U.S. Healthcare Finance Systems. *ASRC Procedia: Global Perspectives in Science and Scholarship*, *1*(01), 1570–1602. https://doi.org/10.63125/x4cta041
[68]. Rifat, C., & Jinnat, A. (2022). Optimization Algorithms for Enhancing High Dimensional Biomedical Data Processing Efficiency. *Review of Applied Science and Technology*, *1*(04), 98–145. https://doi.org/10.63125/2zg6x055
[69]. 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
[70]. Rodríguez-Abreo, O., Rodríguez-Reséndiz, J., Fuentes-Silva, C., Hernández-Alvarado, R., & Falcón, M. D. C. P. T. (2021). Self-tuning neural network PID with dynamic response control. *Ieee Access*, *9*, 65206-65215.
[71]. Roy, S., Menapace, W., Oei, S., Luijten, B., Fini, E., Saltori, C., Huijben, I., Chennakeshava, N., Mento, F., & Sentelli, A. (2020). Deep learning for classification and localization of COVID-19 markers in point-of-care lung ultrasound. *IEEE transactions on medical imaging*, *39*(8), 2676-2687.
[72]. Sai Praveen, K. (2024). AI-Enhanced Data Science Approaches For Optimizing User Engagement In U.S. Digital Marketing Campaigns. *Journal of Sustainable Development and Policy*, *3*(03), 01-43. https://doi.org/10.63125/65ebsn47
[73]. Salleo, C., Grassi, A., & Kyriakopoulos, C. (2020). A comprehensive approach for calculating banking sector risks. *International journal of financial studies*, *8*(4), 69.
[74]. Sayles, J. (2024). Aligning AI Governance with Other Internal Governance Models for Trustworthy AI:"The Convergence of Governance Frameworks". In *Principles of AI Governance and Model Risk Management: Master the Techniques for Ethical and Transparent AI Systems* (pp. 113-172). Springer.
[75]. Schwarz, M., Hinske, L. C., Mansmann, U., & Albashiti, F. (2024). Designing an ml auditing criteria catalog as starting point for the development of a framework. *Ieee Access*, *12*, 39953-39967.
[76]. Shi, X.-J., Qin, Y.-F., & Zhao, L. (2022). Optimal test point placement based on fault diagnosability quantitative evaluation. *Ieee Access*, *10*, 74495-74507.
[77]. Shofiul Azam, T. (2025). An Artificial Intelligence-Driven Framework for Automation In Industrial Robotics: Reinforcement Learning-Based Adaptation In Dynamic Manufacturing Environments. *American Journal of Interdisciplinary Studies*, *6*(3), 38-76. https://doi.org/10.63125/2cr2aq31
[78]. Srokosz, M., Bobyk, A., Ksiezopolski, B., & Wydra, M. (2023). Machine-learning-based scoring system for antifraud CISIRTs in banking environment. *Electronics*, *12*(1), 251.
[79]. Stockinger, K., Bundi, N., Heitz, J., & Breymann, W. (2019). Scalable architecture for Big Data financial analytics: user-defined functions vs. SQL. *Journal of Big Data*, *6*(1), 46.
[80]. Tarrant, C., Colman, A. M., Jenkins, D. R., Chattoe-Brown, E., Perera, N., Mehtar, S., Nakkawita, W. D., Bolscher, M., & Krockow, E. M. (2021). Drivers of broad-spectrum antibiotic overuse across diverse hospital contexts—A qualitative study of prescribers in the UK, Sri Lanka and South Africa. *Antibiotics*, *10*(1), 94.
[81]. Tasnim, K. (2025). Digital Twin–Enabled Optimization of Electrical, Instrumentation, And Control Architectures In Smart Manufacturing And Utility-Scale Systems. *International Journal of Scientific Interdisciplinary Research*, *6*(1), 404–451. https://doi.org/10.63125/pqfdjs15
[82]. Tchapmi, L., Choy, C., Armeni, I., Gwak, J., & Savarese, S. (2017). Segcloud: Semantic segmentation of 3d point clouds. 2017 international conference on 3D vision (3DV),
[83]. Tian, R., Zhang, Y., Yang, L., Zhang, J., Coleman, S., & Kerr, D. (2024). Dynaquadric: dynamic quadric slam for quadric initialization, mapping, and tracking. *IEEE Transactions on Intelligent Transportation Systems*.
[84]. Udayakumar, P. (2022). Designing and Deploying AVD Solution. In *Design and Deploy Microsoft Azure Virtual Desktop: An Essential Guide for Architects and Administrators* (pp. 165-299). Springer.

[85]. Van der Schaar, M., Alaa, A. M., Floto, A., Gimson, A., Scholtes, S., Wood, A., McKinney, E., Jarrett, D., Lio, P., & Ercole, A. (2021). How artificial intelligence and machine learning can help healthcare systems respond to COVID-19. *Machine Learning*, *110*(1), 1-14.

[86]. Van Gansbeke, W., Vandenhende, S., Georgoulis, S., Proesmans, M., & Van Gool, L. (2020). Scan: Learning to classify images without labels. European conference on computer vision,

[87]. Wang, H. E., Donnelly, J. P., Barton, D., & Jarvis, J. L. (2018). Assessing advanced airway management performance in a national cohort of emergency medical services agencies. *Annals of emergency medicine*, *71*(5), 597-607. e593.

[88]. Wang, J., Ding, H., Bidgoli, F. A., Zhou, B., Iribarren, C., Molloi, S., & Baldi, P. (2017). Detecting cardiovascular disease from mammograms with deep learning. *IEEE transactions on medical imaging*, *36*(5), 1172-1181.

[89]. Wang, Y., Li, Y., & Wu, T. (2021). Research on compliance supervision data analysis model based on mass chat records in the inter-bank market. 2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE),

[90]. Yang, X., Song, Z., King, I., & Xu, Z. (2022). A survey on deep semi-supervised learning. *IEEE transactions on knowledge and data engineering*, *35*(9), 8934-8954.

[91]. Yang, Y., Hou, C., Lang, Y., Yue, G., & He, Y. (2019). One-class classification using generative adversarial networks. *Ieee Access*, *7*, 37970-37979.

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

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

[94]. Zaman, M. A. U., Sultana, S., Raju, V., & Rauf, M. A. (2021). Factors İmpacting the Uptake of İnnovative Open and Distance Learning (ODL) Programmes in Teacher Education. *Turkish Online Journal of Qualitative Inquiry*, *12*(6).

[95]. Zehra, S., Faseeha, U., Syed, H. J., Samad, F., Ibrahim, A. O., Abulfaraj, A. W., & Nagmeldin, W. (2023). Machine learning-based anomaly detection in NFV: A comprehensive survey. *Sensors*, *23*(11), 5340.

[96]. Zeydan, E., Arslan, S. S., & Liyanage, M. (2024). Managing distributed machine learning lifecycle for healthcare data in the cloud. *Ieee Access*.

[97]. Zhang, T., & Zhang, W. (2018). Multiple instance learning for credit risk assessment with transaction data. *Knowledge-Based Systems*, *161*, 65-77.