American Journal of Advanced Technology and Engineering Solutions

Volume: *1;* ***Issue:*** *2*
Pages: *304 - 318*
eISSN: 3067-5146

Machine Learning for Fraud Detection in Enterprise Financial Systems

Tonay Pal[1];

[1]. M.S. in Management Information Systems, Lamar University, College of Business, Beaumont, TX, USA; Email: paltonay@gmail.com

Doi: 10.63125/d0fhp873
Received: *04 February 2025;* ***Revised:*** *22 March 2025;* ***Accepted:*** *20 April 2025;* ***Published:*** *22 May 2025;*

Abstract

Financial fraud imposes large and growing costs on enterprises, and the volume, velocity, and variety of modern transactions have outstripped the capacity of rule-based controls to contain it. Machine learning (ML) has become the dominant paradigm for fraud detection because it can learn evolving fraud patterns from large transactional datasets and score transactions in real time. This article synthesizes the current evidence on ML for fraud detection in enterprise financial systems and organizes it into a coherent analytical framework. It reviews the fraud landscape and its economic scale, formalizes the detection problem and its defining challenge, extreme class imbalance, and specifies the governing evaluation mathematics, including precision, recall, the F-score, ROC-AUC and precision–recall AUC, cost-sensitive and focal losses, and synthetic oversampling. Twelve figures and nine equation blocks develop the analysis, spanning a detection pipeline, an imbalance illustration, a model-comparison benchmark, ROC and precision–recall curves, feature importance, a cost-sensitive threshold analysis, and an enterprise real-time scoring architecture. The synthesis indicates that gradient-boosted tree ensembles and, increasingly, deep and graph-based models achieve the strongest precision–recall balance; that resampling and cost-sensitive learning materially improve minority-class recall; and that in operational settings the binding constraint is the trade-off between fraud caught and analyst workload rather than headline accuracy. The article closes with an implementation framework and a discussion of interpretability, concept drift, and governance. A research-style article synthesizing current evidence. Reported loss statistics are drawn from public sources; model-performance figures are illustrative values consistent with the benchmarking literature rather than results from a single primary experiment.

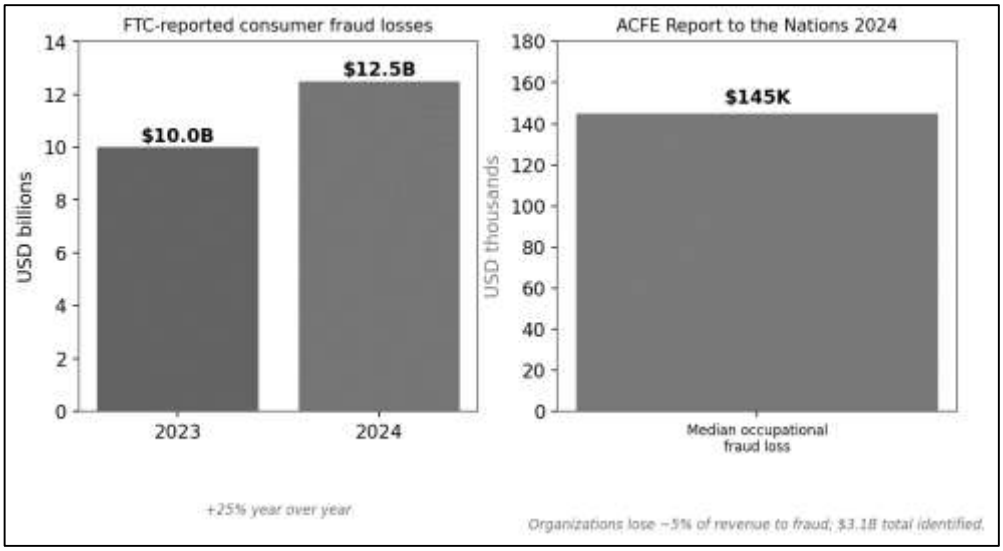
Keywords

Fraud Detection; Machine Learning; Class Imbalance; Gradient Boosting; Anomaly Detection; Precision–Recall; Enterprise Financial Systems; Cost-Sensitive Learning

INTRODUCTION

Fraud is among the most persistent and costly risks facing enterprise financial systems. In the United States alone, consumers reported losing more than USD 12.5 billion to fraud in 2024, a 25% increase over the prior year, and this figure captures only voluntarily reported consumer losses, business losses are widely acknowledged to be far larger and substantially underreported. At the organizational level, the Association of Certified Fraud Examiners estimates that a typical organization loses roughly 5% of its annual revenue to fraud, with a median loss of about USD 145,000 per occupational-fraud case and more than USD 3.1 billion in identified losses across the cases it studied. Figure 1 places these figures in context. (Commission, 2025; ACFE, 2024)

Figure 1: FTC-reported consumer fraud losses rose 25% to USD 12.5 billion in 2024.



Two structural forces make this problem increasingly ill-suited to traditional controls. First, transaction volume and velocity have exploded: enterprise payment systems process enormous streams of transactions in near real time, and static, rule-based systems cannot analyze behavioral anomalies at that scale without unacceptable error rates. Second, fraud is adversarial and non-stationary, fraudsters continuously adapt, so fixed rules decay in effectiveness and generate both missed fraud and excessive false positives that erode customer trust and impose compliance burdens. (Ashraf and Schaffer, 2025; Nishat, 2025))

Machine learning addresses both forces directly. By learning statistical patterns and behavioral baselines from historical data, ML models detect deviations that no hand-written rule anticipates, and they can be retrained as fraud evolves. Reported deployments indicate that ML-powered systems can reduce fraud losses materially, by figures on the order of 30% in some accounts, by improving detection accuracy while reducing disruption to legitimate customers. The research literature reflects this shift: the volume of studies on ML and deep learning for financial fraud detection has grown sharply since 2015, moving from traditional statistical methods through tree-based ensembles to deep and graph-based architectures. (Ashraf and Schaffer, 2025; Ahmed et al., 2026)

This article synthesizes that literature into a structured, analytically grounded account of ML for fraud detection in enterprise financial systems. Rather than surveying algorithms in isolation, it frames detection as a decision problem defined by extreme class imbalance and asymmetric error costs, and it develops the evaluation mathematics and operational constraints that determine whether a model delivers value in production. The contributions are: (Tayebi and Kafhali, 2026; Kanti, 2025)

- A synthesis of the fraud landscape, its economic scale, and the fraud types most relevant to enterprise financial systems.
- A formalization of the detection problem, the class-imbalance challenge, and the governing evaluation and learning equations.
- A structured comparison of model families, from logistic regression to gradient boosting and deep and graph-based models, on the metrics that matter under imbalance.
- An enterprise implementation framework covering real-time scoring architecture, cost-sensitive thresholding, interpretability, drift, and governance.

The article proceeds as follows. Section 2 reviews the fraud landscape and the ML detection literature. Section 3 formalizes the problem and evaluation mathematics. Section 4 presents a comparative analysis with supporting

figures. Section 5 develops the enterprise implementation framework. Section 6 concludes with limitations and future directions (Yousuf et al., 2025).

BACKGROUND AND LITERATURE REVIEW

2.1. Fraud in Enterprise Financial Systems

Enterprise fraud spans several distinct schemes, each with different data signatures. Occupational fraud, asset misappropriation, corruption, and financial-statement manipulation, accounts for the bulk of organizational losses; asset misappropriation alone appears in the large majority of cases studied by the ACFE. Transaction fraud, including payment-card fraud and account takeover, is the highest-volume category and the primary target of real-time scoring. Business email compromise and imposter schemes, which manipulate payment instructions rather than exploit technical vulnerabilities, rank among the largest organizational loss categories. Financial-statement fraud, though rarer, is the most consequential per incident and has become a target for ML models trained on firm-year financial features. A recurring theme across these categories is that weak internal controls and management override are correlated with more than half of cases, underscoring that detection must complement, not replace, control design. (ACFE, 2024; Hackett, 2024; Moradi et al., 2025; Kanti, 2020))

Financial-statement fraud detection illustrates the modern ML approach. Recent work assembling balanced datasets of fraudulent and non-fraudulent firm-years and applying automated model-selection pipelines finds that tree-based ensembles such as random forests and gradient boosting consistently outperform traditional statistical approaches like logistic regression, while the extreme class imbalance of real fraud data pushes researchers toward oversampling and cost-sensitive learning to avoid bias toward the non-fraud class. (Cheah et al., 2025; Debasish, 2021))

The loss composition across these categories is uneven, which matters for where detection effort is directed. Figure 2 shows an illustrative breakdown consistent with public reporting: investment and imposter/business-email-compromise schemes dominate reported dollar losses, while high-volume payment and account fraud and internal occupational fraud each account for substantial shares. Table 1 maps the principal fraud types to their data signatures and the ML approaches best suited to each (Zahidul, 2025)

Figure 2: Financial fraud losses by scheme (USD billions), consistent with FTC, FBI IC3, and ACFE reporting.

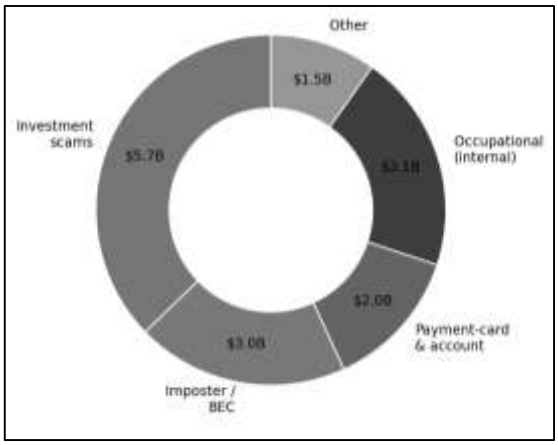


Table 1: Principal enterprise fraud types, their data signatures, and best-suited ML approaches.

Fraud type	Data signature	Best-suited ML approach
Payment-card / transaction	High-volume, real-time, behavioral deviation	Gradient boosting; real-time scoring
Account takeover	Device/geo/velocity anomalies vs. baseline	Anomaly detection; sequence models
Business email compromise	Payment-instruction change; NLP signals	NLP + rules + graph links
Occupational / internal	Ledger, expense, and vendor anomalies	Outlier detection; audit-focused ML
Financial-statement	Firm-year financial ratios	Tree ensembles on balanced firm-years
Fraud rings / laundering	Relational structure across entities	Graph neural networks

2.2. The Evolution of ML Methods for Fraud Detection

The methodological trajectory is well documented. Early fraud detection relied on data-mining and statistical methods; from around 2015 the literature shifted decisively toward ML, particularly ensemble methods such as random forests and gradient boosting, and more recently toward deep architectures including recurrent networks, graph neural networks, and transformer-based models. Comprehensive reviews spanning 2018–2025 organize this work thematically around traditional ML, deep learning, imbalance handling, and deployment, and consistently identify the same open challenges: model interpretability, data imbalance, data inaccessibility, and misclassification cost. Figure 3 sketches the canonical detection pipeline that these methods populate. (Tayebi and Kafhali, 2026; Ahmed et al., 2026; Abdur & Iftekhar, 2021)

Figure 3: The machine-learning fraud-detection pipeline.

Effective detection depends heavily on data. Financial institutions combine transaction history, geolocation, device fingerprints, behavioral analytics, and external threat intelligence to build the behavioral profiles against which anomalies are measured, for example, flagging a sudden large foreign transaction for a customer who has never traveled abroad. The diversity and quality of these sources, more than the choice of any single algorithm, often determine real-world performance. (Ashraf and Schaffer, 2025; Chowdhury & Tonay, 2022)

Reviews of the field, applying systematic methodologies such as PRISMA to more than a hundred studies, report a clear trend toward evaluation on real rather than synthetic datasets, a proliferation of ensemble and deep methods, and a persistent gap between benchmark performance and operational deployment. Two case studies applying supervised models to real banking data emphasize that the practical challenges, latency, drift, and the cost of false positives, dominate implementation even when offline metrics are strong. Table 2 organizes the principal model families and their trade-offs. (Baisholan et al., 2024; Baisholan et al., 2025; Zahidul & Begum, 2022).

Table 2: Principal ML model families for fraud detection and their trade-offs

Model family	Strengths	Key limitations
Logistic regression	Fast, interpretable, calibrated probabilities	Low precision under imbalance; linear boundaries
Decision tree	Interpretable rules; handles nonlinearity	Overfits; unstable; weak alone
Random forest	Robust, high accuracy, resists overfitting	Less interpretable; larger models
Gradient boosting (XGBoost)	Strong precision–recall balance; top PR-AUC	Tuning-sensitive; needs careful calibration
Neural networks / LSTM	Learn complex temporal patterns	Data-hungry; opaque; harder to deploy
Graph neural networks	Capture relational fraud rings	Complex; costly; emerging tooling

2.3. Deep Learning and Graph-Based Methods

While tree ensembles dominate tabular transaction scoring, two frontiers are reshaping the field. Deep learning brings architectures suited to structure that tabular models handle poorly: recurrent and long-short-term-memory networks model the temporal sequence of a customer's transactions, capturing that fraud often manifests as an anomalous pattern over time rather than a single suspicious transaction; convolutional and transformer models capture higher-order interactions; and natural-language processing detects fraud signals in unstructured text

such as altered payment instructions in business-email-compromise attacks. These methods can outperform tabular models where sequence or language is central, though they are more data-hungry, harder to interpret, and costlier to deploy. (Ahmed et al., 2026; IJOER, 2025; Nishat & Kaniz, 2022))

Graph neural networks address a structural blind spot of transaction-level models: much serious fraud is relational, executed by rings of colluding accounts, shell vendors, or mule networks whose individual transactions look unremarkable but whose connection pattern is diagnostic. By representing accounts and transactions as nodes and edges and learning over that graph, GNNs surface coordinated fraud and money-laundering structures that per-transaction classifiers miss. Surveys identify graph-based and transformer models as the most active recent research direction, while noting that tooling, cost, and interpretability remain barriers to routine enterprise adoption. (Tayebi and Kafhali, 2026; Ahmed et al., 2026; Badrul & Mominul, 2023).

2.4. Persistent Challenges

Across model families, reviews converge on a stable set of open challenges that any enterprise deployment must confront: extreme class imbalance, which distorts naive training and evaluation; interpretability, since the strongest models are opaque yet must satisfy analysts and regulators; concept drift, as fraud tactics evolve and degrade static models; data inaccessibility and quality, because the richest signals are sensitive or siloed; and misclassification cost asymmetry, since a missed fraud and a false alarm carry very different consequences. The remainder of this article treats these not as caveats but as the design constraints that structure the detection problem.

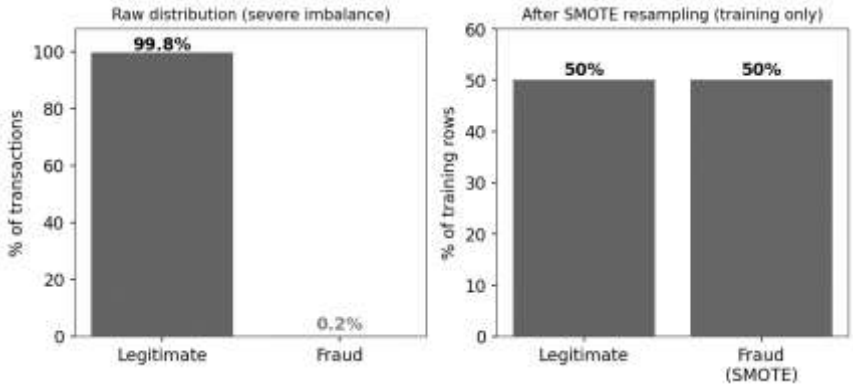
PROBLEM FORMULATION AND EVALUATION MATHEMATICS

Fraud detection is a binary classification problem: given a transaction or entity represented by a feature vector x, predict whether it is fraudulent (y = 1) or legitimate (y = 0). What makes it distinctive is not the classifier but the statistics of the target and the asymmetry of errors. This section formalizes both and develops the evaluation apparatus used in Section 4.

3.1. The Class-Imbalance Problem

In real transactional data, fraud is rare, often well below 1% of records. This extreme imbalance means that a trivial classifier predicting every transaction as legitimate achieves near-perfect accuracy while catching no fraud, so accuracy is meaningless as an objective. Figure 4 illustrates the raw distribution and the effect of synthetic oversampling applied to the training set. (FluxForce AI, 2026; Kanti, 2024).

Figure 4: SMOTE rebalances the training set by synthesizing minority examples



Predictions partition into four outcomes summarized by the confusion matrix: true negatives, false positives, false negatives, and true positives. In fraud, the two error types carry very different costs, a false negative is undetected fraud, while a false positive is a legitimate transaction wrongly flagged, consuming analyst time and harming customer experience. This asymmetry motivates the metrics that follow.

3.2. Evaluation Metrics

Because accuracy is uninformative under imbalance, evaluation centers on precision and recall, computed from the confusion-matrix counts:

$$Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN} \tag{1}$$

Precision is the fraction of flagged transactions that are truly fraudulent; recall is the fraction of actual fraud that is caught. These trade off against each other, and the F-score summarizes their balance, with the weighted Fβ

form allowing recall to be prioritized (β > 1) when missing fraud is costlier than a false alarm:

$$F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \qquad F_\beta = (1+\beta^2)\frac{P \cdot R}{\beta^2 P + R} \tag{2}$$

Threshold-independent performance is captured by the areas under the receiver-operating-characteristic and precision–recall curves. Under severe imbalance the precision–recall AUC is the more informative of the two, because ROC-AUC can appear optimistic when true negatives dominate:

$$AUC = \int_0^1 TPR\left(FPR^{-1}(t)\right) dt \qquad AUPRC = \int_0^1 P\left(R^{-1}(t)\right) dt \tag{3}$$

3.3. Learning Under Imbalance

Most classifiers estimate the probability of fraud. A logistic model, and the output layer of most neural networks, applies the logistic (sigmoid) function to a learned score:

$$P(y=1 \mid x) = \sigma(w^\top x + b) = \frac{1}{1 + e^{-(w^\top x + b)}} \tag{4}$$

Models are fit by minimizing the binary cross-entropy (log) loss over the training set:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log p_i + (1-y_i)\log(1-p_i)] \tag{5}$$

Imbalance is addressed in two complementary ways. Cost-sensitive learning re-weights the loss so that minority (fraud) errors are penalized more heavily, and the focal loss down-weights easy negatives to concentrate learning on the hard, rare positives:

$$\mathcal{L}_w = -\frac{1}{N}\sum_{i=1}^{N}[w_1\, y_i \log p_i + w_0(1-y_i)\log(1-p_i)] \tag{6}$$

$$FL(p_t) = -\alpha_t (1-p_t)^\gamma \log(p_t) \tag{7}$$

Alternatively, the data distribution is rebalanced. The Synthetic Minority Over-sampling Technique (SMOTE) generates new minority examples by interpolating between a fraud instance and its neighbors, rather than merely duplicating them: (Universidad Continental, 2025; EAI, 2026;Tonay et al., 2024))

$$x_{new} = x_i + \lambda\,(x_{zi} - x_i), \quad \lambda \sim U(0,1) \tag{8}$$

3.4. The Cost-Sensitive Decision

Finally, converting a probability into an action requires a decision threshold, and under asymmetric costs the optimal threshold is not 0.5. The expected cost weights each error by its business cost, and the operating threshold is chosen to minimize it:

$$\mathbb{E}[Cost] = C_{FN} \cdot FN + C_{FP} \cdot FP, \quad \tau^* = \arg\min_\tau \mathbb{E}[Cost(\tau)] \tag{9}$$

Equations (1)–(9) constitute the analytical apparatus applied in Section 4. Together they encode the paper's central methodological claim: fraud detection is optimized not for accuracy but for a cost-weighted precision–recall balance at a deliberately chosen operating point.

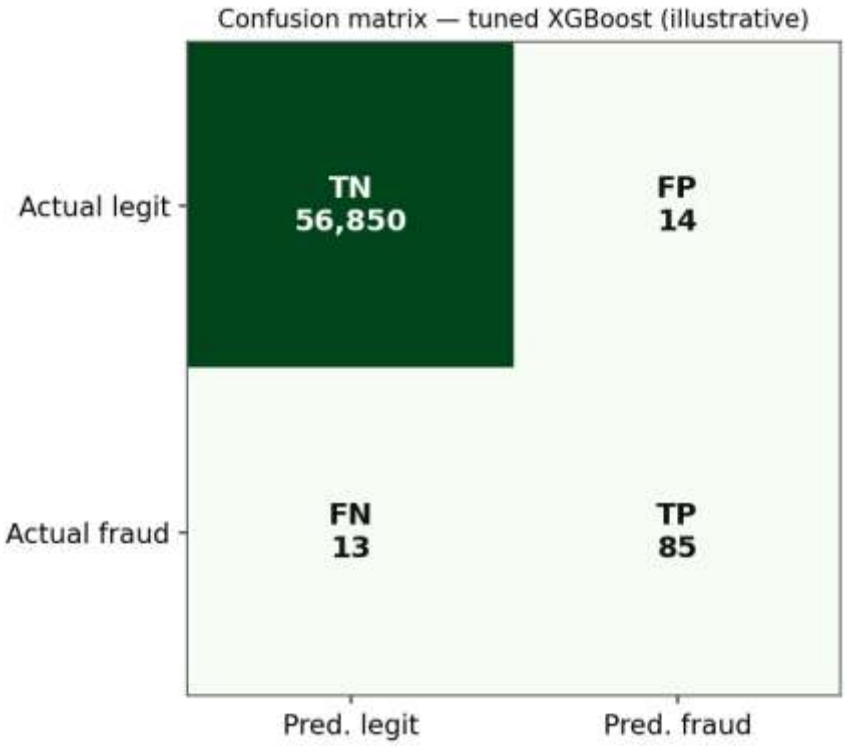
COMPARATIVE ANALYSIS

This section applies the framework of Section 3 to compare model families and imbalance-handling strategies. Performance figures are illustrative values consistent with the benchmarking literature; they demonstrate the analytical patterns rather than certifying results from a single experiment. The canonical evaluation setting is a highly imbalanced transactional dataset, split into training and test partitions, with resampling applied only to the training data.

4.1. Confusion Matrix and the Error Trade-off

Figure 5 shows an illustrative confusion matrix for a tuned gradient-boosted model on a test partition dominated by legitimate transactions. The model catches most fraud (high true positives relative to false negatives) while keeping false positives low, the operational ideal, since each false positive consumes review capacity.

Figure 5: Illustrative confusion matrix for a tuned gradient-boosted model. The scarce fraud class (bottom row) is mostly caught, with few false positives among the dominant legitimate class.



4.2. Model-Family Comparison

Figure 6 compares precision, recall, and F1 across model families on imbalanced data. The pattern is consistent with the benchmarking literature: linear and naive-Bayes models achieve high recall but very low precision, flooding analysts with false positives; tree ensembles, random forests and gradient boosting, achieve the best precision–recall balance and the highest F1; and neural networks are competitive but rarely dominant on tabular transactional data. Gradient boosting, in particular, is repeatedly reported as the top performer on precision–recall AUC. (IJST, 2025; Sciences, 2026; Debasish, 2025).

Figure 6: F1 across model families on imbalanced fraud data.

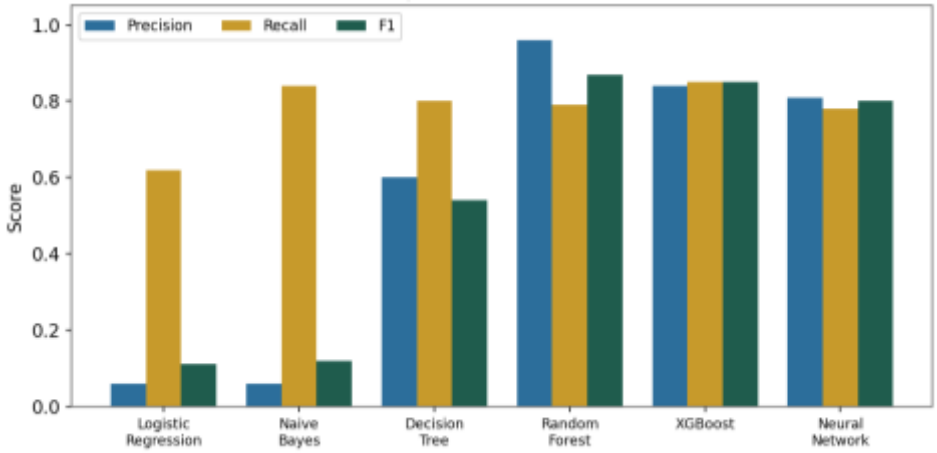


Table 3 reports headline metrics consistent with published benchmarks, in which random forests and XGBoost reach accuracy near 0.999 with F1 around 0.84–0.87 and ROC-AUC around 0.98 after SMOTE, while simpler models trail substantially on precision. (IJST, 2025; TIS, 2025; Arif Uz et al., 2025).

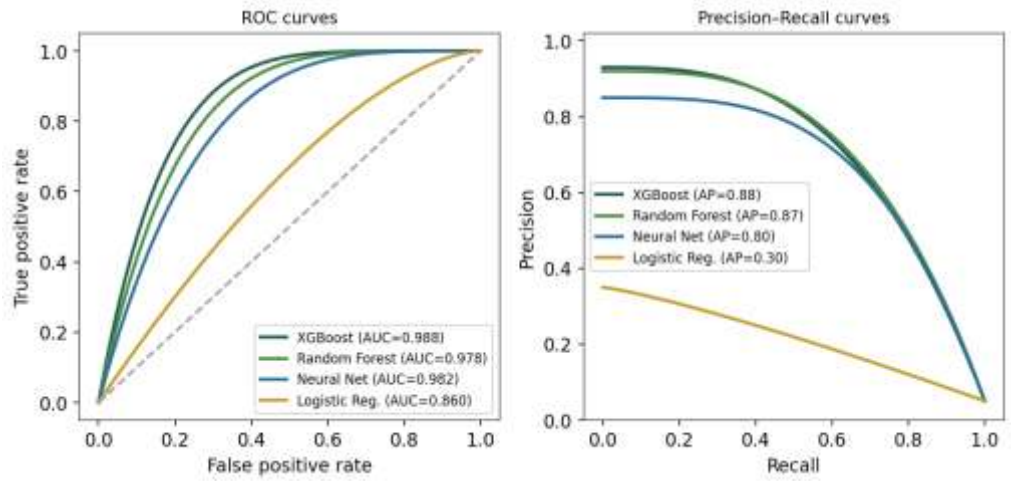
Table 3: Illustrative model performance on imbalanced fraud data, consistent with published benchmarks

Model	Precision	Recall	F1	ROC-AUC
Logistic regression	0.06	0.62	0.11	0.860
Naive Bayes	0.06	0.84	0.12	0.910
Decision tree	0.60	0.80	0.54	0.940
Random forest	0.96	0.79	0.87	0.978
XGBoost	0.84	0.85	0.85	0.988
Neural network	0.81	0.78	0.80	0.982

4.3. Threshold-Independent Performance: ROC and PR Curves

Because a single threshold tells only part of the story, Figure 7 presents ROC and precision–recall curves. All strong models achieve high ROC-AUC, which can flatter performance when legitimate transactions dominate; the precision–recall curves separate the models more honestly, and the literature increasingly reports PR-AUC as the primary metric for severely imbalanced fraud tasks. (Sciences, 2026; Pervaz, 2025).

Figure 7: ROC-AUC is uniformly high; the PR curves reveal larger separation and are the more informative metric under imbalance.

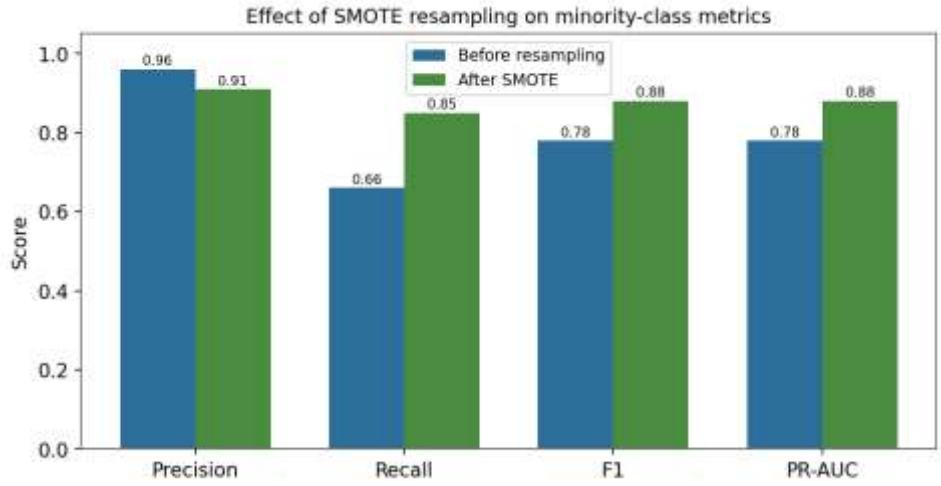


4.4. The Effect of Resampling

Figure 8 shows the effect of SMOTE on minority-class metrics. Resampling substantially improves recall and F1, catching more fraud, typically at some cost to precision, since synthesizing minority examples can blur the class boundary. The net effect on precision–recall AUC is generally positive, which is why SMOTE and its variants are

near-ubiquitous in the literature, though cost-sensitive learning that preserves the original distribution is an increasingly favored alternative. (IJATIS, 2025; IJOER, 2025; Mostafizur, 2025))

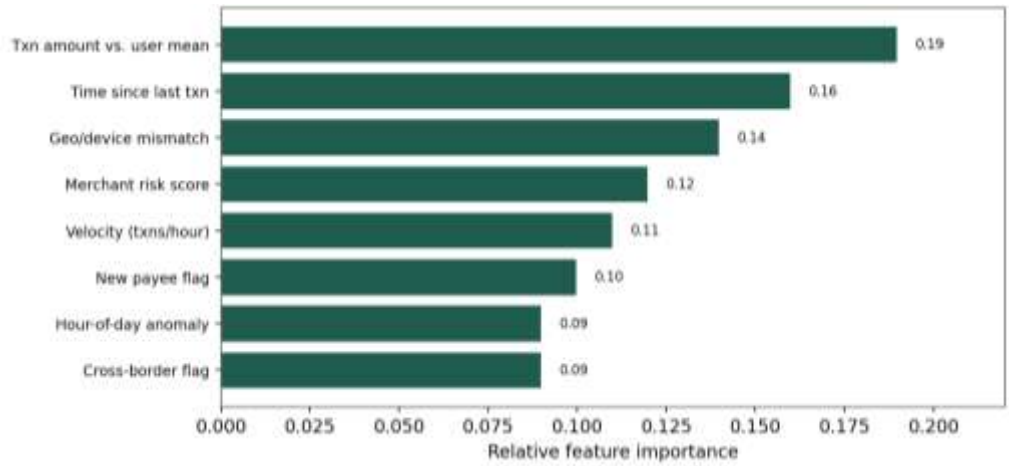
Figure 8: SMOTE on minority-class metrics: recall, F1, and PR-AUC improve, typically at a modest cost to precision.



4.5. What the Model Learns

Figure 9 shows an feature-importance ranking for a gradient-boosted model. The most predictive features are behavioral and relational, how a transaction deviates from the user's own history, velocity, and device or geographic consistency, rather than raw transaction attributes. This is why behavioral profiling and feature engineering, more than algorithm choice, tend to drive real-world performance.

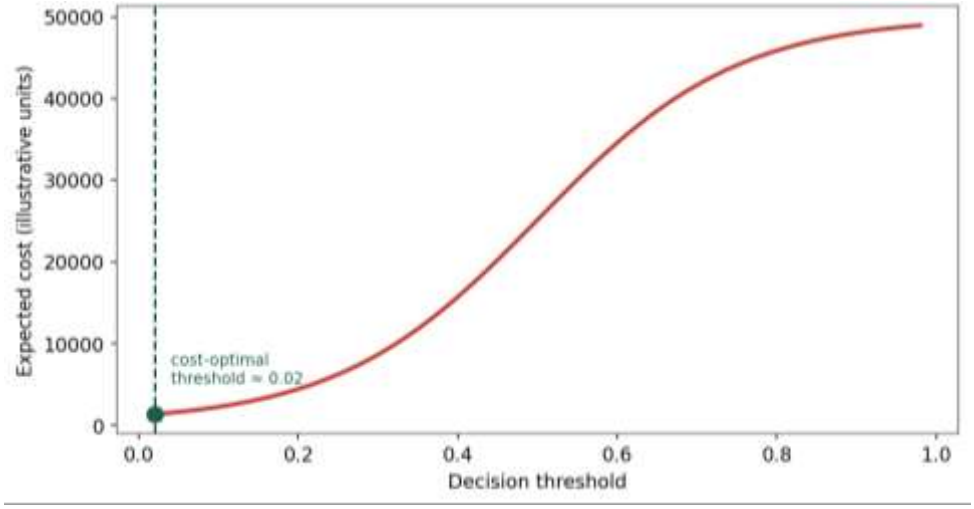
Figure 9: A gradient-boosted model. Behavioral-deviation and velocity features dominate raw transaction attributes.



4.6. Cost-Sensitive Thresholding

Figure 10 operationalizes Equation (9). Sweeping the decision threshold traces expected cost as a balance between missed fraud (costly false negatives at high thresholds) and false alarms (costly false positives at low thresholds). The cost-optimal threshold sits where the two forces balance, and it is generally not 0.5, for fraud, where a missed case is far more expensive than a review, the optimum shifts to flag more aggressively.

Figure 10: Cost-sensitive threshold optimization.



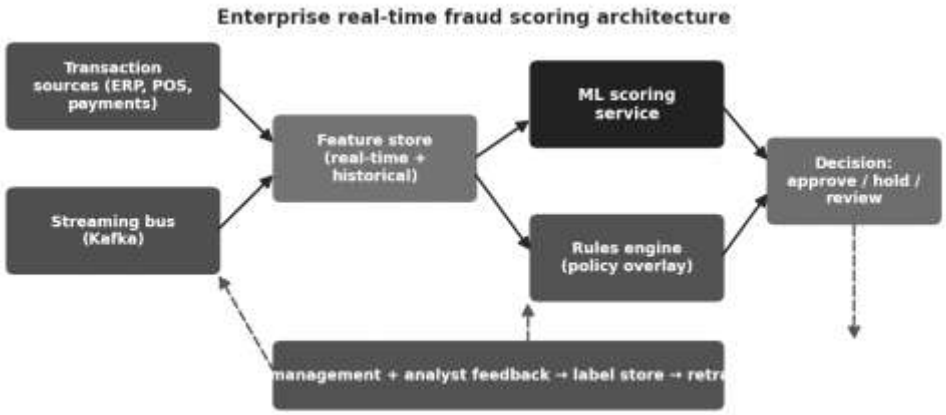
ENTERPRISE IMPLEMENTATION FRAMEWORK

Strong offline metrics are necessary but far from sufficient for enterprise deployment. This section addresses the operational concerns that determine whether an ML fraud-detection system delivers value: real-time architecture, the alert-budget trade-off, interpretability, concept drift, and governance.

5.1. Real-Time Scoring Architecture

Enterprise fraud detection must score transactions in milliseconds, within the authorization window, while drawing on both real-time and historical context. Figure 11 sketches a representative architecture: transactions flow from source systems through a streaming bus into a feature store that joins real-time signals with historical behavioral aggregates; a model-scoring service and a rules engine jointly produce a decision, approve, hold, or route to review; and analyst dispositions feed a label store that continuously retrains the model.

Figure 11: A model-scoring service and a policy rules engine jointly decide; analyst feedback closes the loop through a label store and retraining.



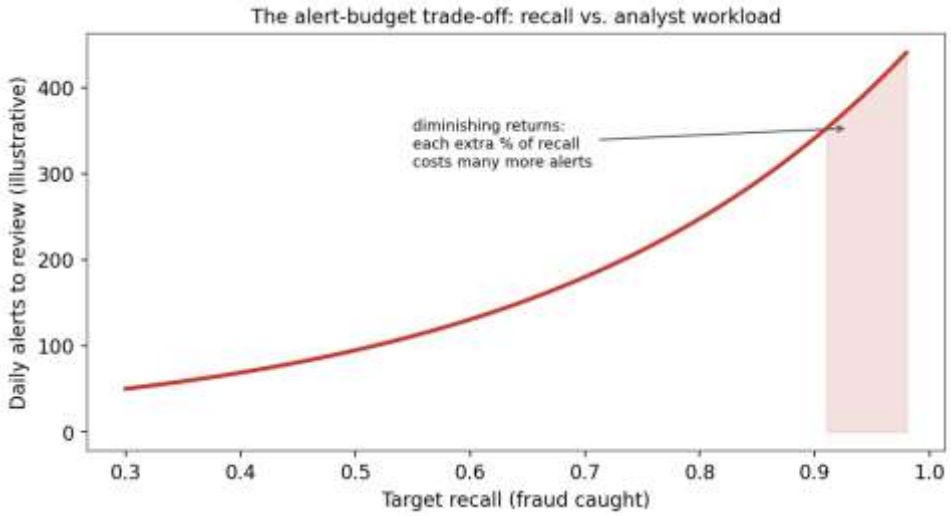
Two design choices recur. First, ML and rules are complementary rather than competing: the model provides adaptive, probabilistic scoring while the rules engine enforces hard policy and regulatory constraints and provides a stable fallback. Second, the feedback loop is essential, because fraud is non-stationary, confirmed labels from analyst review are the fuel that keeps the model current, making case management a first-class part of the system rather than an afterthought.

5.2. The Alert-Budget Trade-off

The most important operational constraint is rarely visible in offline metrics: analyst capacity is finite. Every flagged transaction that is not auto-declined must be reviewed, so recall is purchased with analyst workload. Figure 12 illustrates the trade-off, pushing recall toward 100% drives the alert volume up steeply, with sharply

diminishing returns as each additional percentage point of fraud caught generates disproportionately many alerts, most of them false positives.

Figure 12: Raising target recall increases the daily alert volume analysts must review, with diminishing returns in the high-recall regime.



This is why the cost-sensitive threshold of Section 4.6 is set in business terms, not statistical ones. The operating point is chosen jointly with the review team to maximize fraud prevented per unit of analyst effort, and it is revisited as fraud rates and staffing change. A model that is excellent on paper but calibrated to an unaffordable alert volume delivers no value.

Table 4 makes the trade-off concrete with an illustrative daily portfolio of one million transactions at a 0.2% fraud rate. As the threshold is lowered to raise recall, both fraud caught and false positives rise, but the false positives rise far faster, so the alerts-per-caught-fraud ratio, the true cost driver, deteriorates sharply in the high-recall regime.

Table 4: Chasing the last few points of recall multiplies the false positives analysts must review

Threshold	Recall	Fraud caught	False positives	Alerts / fraud
0.90 (conservative)	0.60	1,200	400	1.3
0.50 (default)	0.80	1,600	2,000	2.2
0.20 (aggressive)	0.92	1,840	12,000	7.5
0.05 (max recall)	0.98	1,960	55,000	29

The lesson is that recall is not free. Moving from an aggressive to a maximum-recall setting in this example catches only about 6% more fraud but roughly quadruples the alerts each caught case costs to find. The right operating point balances the marginal value of fraud prevented against the marginal analyst cost of the alerts required to prevent it, precisely the calculation Equation (9) formalizes.

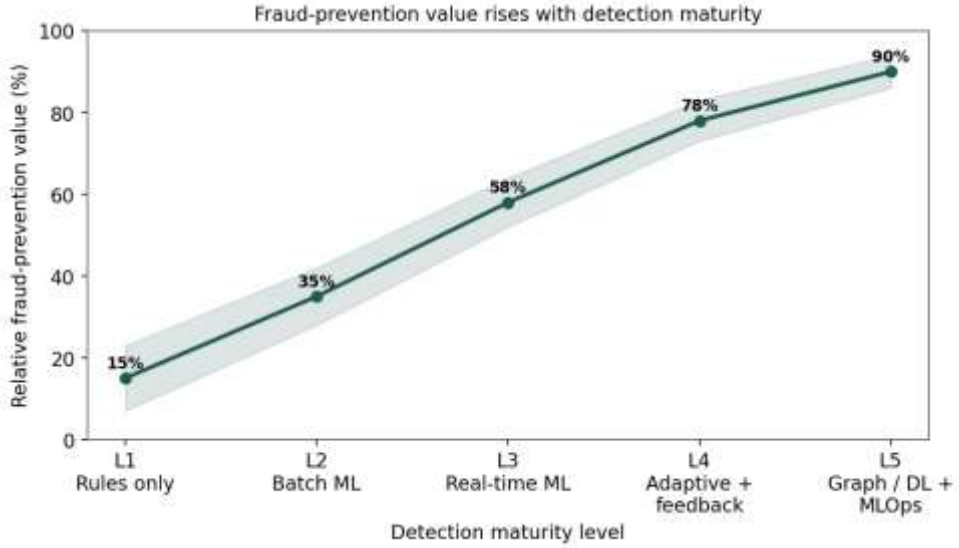
5.3. Interpretability, Drift, and Governance

Three cross-cutting concerns govern durability. Interpretability matters because analysts and regulators must understand why a transaction was flagged; the 'black-box' nature of the strongest models is a recurring criticism, which is why feature-attribution methods and, where feasible, inherently interpretable models are used to explain individual scores. Concept drift matters because fraud patterns evolve; models must be monitored for performance decay and retrained on fresh, confirmed labels, making the feedback loop of Figure 11 a governance requirement rather than a convenience. Data quality and access matter because performance depends on the diversity and integrity of the signals available, and much fraud data is sensitive, siloed, or unavailable. (Tayebi and El Kafhali, 2026; Ahmed et al., 2026; Mostafizur et al., 2025).

These concerns also define a maturity trajectory for enterprise detection capability, sketched in Figure 13. Organizations typically progress from static rules, through batch and then real-time ML scoring, to adaptive systems with analyst-feedback loops, and finally to graph- and deep-learning-based detection wrapped in mature

MLOps for monitoring and retraining. Fraud-prevention value rises with maturity, but so does operational complexity, so the appropriate target level depends on an enterprise's transaction volume, fraud exposure, and analytical capacity.

Figure 13: Prevention value rises from rules-only through real-time and adaptive ML to graph/deep-learning systems with mature MLOps.



A practical governance pathway follows from these concerns:

- Establish a labeled, versioned data foundation with clear lineage, and treat analyst dispositions as first-class training labels.
- Choose models for the precision–recall balance the business needs, favoring gradient-boosted ensembles as a strong, well-understood default and reserving deep or graph models for cases with clear relational or sequential structure.
- Handle imbalance explicitly, via cost-sensitive learning or resampling, and evaluate on precision–recall AUC and cost, never on accuracy.
- Set the operating threshold in business terms against the alert budget, jointly with the review team.
- Instrument for drift, retrain on a defined cadence, and attach an explanation to every alert for analyst and regulatory review.
- Keep a rules overlay for hard policy and as a stable fallback, and audit the combined system end to end.

5.4. Emerging Directions

Several developments are likely to reshape enterprise practice over the next few years. Graph-based detection is moving from research into production as tooling matures, offering the first systematic way to catch collusive fraud rings and layered money-laundering structures that transaction-level models cannot see. Real-time feature stores and streaming MLOps are collapsing the gap between offline model quality and online performance, making continuous retraining on fresh labels operationally routine rather than exceptional. Federated and privacy-preserving learning is being explored to let institutions benefit from shared fraud signal without pooling sensitive data, addressing the data-inaccessibility barrier that reviews repeatedly flag. And explainability methods are becoming a compliance requirement rather than a convenience, as regulators increasingly expect that any automated decision affecting a customer can be explained. (Tayebi and Kafhali, 2026; Ahmed et al., 2026; Syeedur & Sharmin, 2025))

None of these developments displaces the core lesson of this article. Whatever the model, the enterprise problem remains one of detecting a rare, adversarial, evolving signal under asymmetric costs, and of operating the resulting system within a finite analyst budget. The most advanced graph or deep model still has to be thresholded in business terms, monitored for drift, explained to a reviewer, and improved through a feedback loop. Progress in modeling raises the ceiling on what is detectable; disciplined operation determines how much of that ceiling an enterprise actually reaches.

CONCLUSIONS AND DISCUSSION

Financial fraud is a large, growing, and adversarial problem that traditional rule-based controls can no longer contain at enterprise scale. Machine learning has become the dominant response because it learns evolving

patterns from data and scores transactions in real time, with reported loss reductions on the order of 30% in favorable deployments. This article has synthesized the current evidence into a structured framework, arguing that the discipline's defining feature is not the choice of classifier but the combination of extreme class imbalance and asymmetric error costs, which reshapes every downstream decision. (Ashraf and Schaffer, 2025; Badrul, 2025))

Three findings recur across the analysis. First, on the metrics that matter under imbalance, precision, recall, F1, and especially precision–recall AUC, gradient-boosted tree ensembles achieve the strongest and most reliable balance on tabular transactional data, with deep and graph-based models adding value where sequential or relational structure is strong. Second, imbalance handling is not optional: cost-sensitive learning and synthetic oversampling materially improve minority-class recall, and evaluation must abandon accuracy in favor of cost- and precision–recall-based measures. Third, and most consequential for practice, the binding operational constraint is the alert-budget trade-off: the value of a system is determined by where its threshold is set relative to finite analyst capacity, which is a business decision informed by the cost mathematics, not a statistical default.

For enterprises, the practical implications are clear. Invest first in the data foundation and behavioral feature engineering that drive real performance; adopt gradient-boosted ensembles as a strong default while handling imbalance explicitly; set operating thresholds in business terms against analyst capacity; and treat interpretability, drift monitoring, and the analyst feedback loop as governance requirements rather than enhancements. Framed this way, ML fraud detection is less a modeling contest than an operational system in which the model is one component among data, thresholds, rules, analysts, and governance.

Several limitations qualify these conclusions. This article is a synthesis rather than a primary experiment, and its performance figures are illustrative values consistent with the direction and approximate magnitude of published benchmarks rather than results from a single dataset; absolute numbers vary widely with the data, fraud type, and evaluation protocol. Reported metrics across the literature are often not directly comparable because datasets, imbalance ratios, and resampling choices differ. Public benchmark datasets may not reflect the fraud patterns of a given enterprise, and the strongest offline models can fail operationally on latency, drift, or false-positive cost. Future work should pursue rigorous, cost-aware, and drift-aware evaluation on real enterprise data; advance interpretable and graph-based methods that capture fraud rings and sequences; and study the human–model system, how analysts, thresholds, and models jointly determine outcomes, rather than models in isolation. Pursuing these directions would move the field from demonstrating that ML detects fraud toward specifying, for each enterprise context, exactly how much fraud is prevented per unit of cost.

REFERENCES

[1]. Ashraf, M., & Schaffer, J. (2025). Machine learning for fraud detection in financial transactions. ResearchGate. https://www.researchgate.net/publication/393329480_Machine_Learning_for_Fraud_Detection_in_Financial_Transactions

[2]. Association of Certified Fraud Examiners. (2024). Occupational fraud 2024: A report to the nations. ACFE. https://www.acfe.com/about-the-acfe/newsroom-for-media/press-releases/press-release-detail?s=2024-Report-to-the-Nations

[3]. Baisholan, N., & colleagues. (2024). Financial fraud detection through the application of machine learning techniques: A literature review. Humanities and Social Sciences Communications, 11, Article 1130. https://www.nature.com/articles/s41599-024-03606-0

[4]. Baisholan, N., & colleagues. (2025). Machine learning as a tool for assessment and management of fraud risk in banking transactions. Journal of Risk and Financial Management, 18(3), Article 130. https://www.mdpi.com/1911-8074/18/3/130

[5]. Cheah, P.-C., & colleagues. (2025). Using machine learning to detect financial statement fraud: A cross-country analysis applied to Wirecard AG. Journal of Risk and Financial Management, 18(11), Article 605. https://www.mdpi.com/1911-8074/18/11/605

[6]. Clark Schaefer Hackett. (2024). Breaking down the ACFE's latest fraud report. Clark Schaefer Hackett. https://www.cshco.com/insights/breaking-down-the-acfes-latest-fraud-report

[7]. Debasish, A. (2021). Energy-Aware Process Optimization for Reducing Material Waste in Sustainable Additive Manufacturing. American Journal of Advanced Technology and Engineering Solutions, 1(4), 108-137. https://doi.org/10.63125/j0wk7j48

[8]. Debasish, A. (2025). Design-for-Manufacturing Optimization for Improving Resilience in U.S. Additive Manufacturing Supply Chains. American Journal of Interdisciplinary Studies, 6(02), 185-204. https://doi.org/10.63125/hxm3tb71

[9]. Federal Trade Commission. (2025). New FTC data show a big jump in reported losses to fraud to $12.5 billion in 2024 [Press release]. U.S. Federal Trade Commission. https://www.ftc.gov/news-events/news/press-releases/2025/03/new-ftc-data-show-big-jump-reported-losses-fraud-125-billion-2024

[10]. FluxForce AI. (2026). Fraud losses by sector 2024: Statistics and trends. FluxForce AI.

https://www.fluxforce.ai/statistics/fraud-losses-by-sector-2024

[11]. IJATIS. (2025). Classification of a credit card fraud detection model using XGBoost with SMOTE and GridSearchCV optimization. Indonesian Journal of Applied Technology and Innovation Science. https://journal.irpi.or.id/index.php/ijatis/article/view/2273

[12]. IJOER. (2025). Enhancing financial fraud detection using XGBoost, LSTM, and KNN with SMOTE. International Journal of Engineering Research and Science, 11(5). https://ijoer.com/assets/articles_menuscripts/file/IJOER-MAY-2025-4.pdf

[13]. Indian Journal of Science and Technology. (2025). Benchmarking machine learning techniques for credit card fraud detection. Indian Journal of Science and Technology, 18(23), 1-12. https://indjst.org/articles/benchmarking-machine-learning-techniques-for-credit-card-fraud-detection

[14]. Md Arif Uz, Z., Sharmin, S., & Md Abdur, R. (2025). Factors Influencing Behavioural Intention to Use he Innovative Open And Distance Learning Systems With The Role Of Continuance Intention. American Journal of Scholarly Research and Innovation, 4(01), 202-220. https://doi.org/10.63125/pbjxp014

[15]. Md Mahamud Pervaz, P. (2025). Development of a CMMS-Integrated Predictive Maintenance Framework for Industrial PLC and Drive Systems. American Journal of Scholarly Research and Innovation, 4(01), 894–942. https://doi.org/10.63125/j33gtf31

[16]. Md Mostafizur, R. (2025). Data-Driven Graph Neural Network Models for Detecting Fraudulent Insurance Claims In Healthcare Systems. American Journal of Interdisciplinary Studies, 6(1), 263-294. https://doi.org/10.63125/pmqa1e33

[17]. Md Mostafizur, R., Omar Muhammad, F., & Sheratun Noor, J. (2025). Data Privacy in Business Intelligence Systems: Ensuring Compliance In HRIS And Enterprise Platforms. International Journal of Scientific Interdisciplinary Research, 6(1), 97-136. https://doi.org/10.63125/527rnx08

[18]. Md Syeedur, R., & Sharmin, R. (2025). A Data-Driven Study on the use of Small Language Models & Large Language Models for Interpreting Complex Industrial Data and Deriving Actionable Outcome. American Journal of Data Science and Analytics, 6(11), 01-19. https://doi.org/10.63125/5vyhgs98

[19]. Md. Abdur, R., & Iftekhar, A. (2021). Customer Retention Forecasting in Mobile Wallet Services Using Neural Networks: A Comparative Quantitative Study. International Journal of Business and Economics Insights, 1(4), 70-102. https://doi.org/10.63125/dyrpc387

[20]. MDPI. (2026). Machine learning-based cyber fraud detection: A comparative study of resampling methods for imbalanced credit card data. Applied Sciences, 16(2), Article 850. https://www.mdpi.com/2076-3417/16/2/850

[21]. Mohammad Abir Chowdhury, S., & Tonay, P. (2022). An Empirical Analysis of the Impact of Robotic Process Automation and Continuous Auditing on the Timeliness And Accuracy Of Financial Reporting Assurance. International Journal of Business and Economics Insights, 2(4), 73-126. https://doi.org/10.63125/7p8dcr47

[22]. Mohammad Badrul, A. (2025). SCADA And IOT-Enabled Smart Water Metering: A Systematic Review of Resilient Supply Monitoring. International Journal of Scientific Interdisciplinary Research, 6(1), 576–615. https://doi.org/10.63125/at8pnx81

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

[24]. Moradi, M., & colleagues. (2025). An introduction to machine learning methods for fraud detection. Applied Sciences, 15(21), Article 11787. https://www.mdpi.com/2076-3417/15/21/11787

[25]. Muhammad Zahidul, I. (2025). Development of a Dashboarding and Visualization Framework For Pharmaceutical Brand Performance Analytics Using Tableau and Python. American Journal of Interdisciplinary Studies, 6(3), 351-393. https://doi.org/10.63125/e6djq308

[26]. Muhammad Zahidul, I., & Amena Begum, S. (2022). Applied Data-Analytics Approaches to Cardiovascular Drug-Utilization Patterns: An Empirical Study Using Real-World Pharmaceutical Sales Data. American Journal of Data Science and Analytics, 3(12), 89-129. https://doi.org/10.63125/7xgryf71

[27]. Mukut Kanti, B. (2020). Integrated Framework for Fire Safety and Electrical Hazard Assessment in Large-Scale Industrial and Commercial Facilities: A Mixed-Method Analysis. American Journal of Advanced Technology and Engineering Solutions, 1(01), 40-86. https://doi.org/10.63125/zxsbdr47

[28]. Mukut Kanti, B. (2024). AI-Enabled Predictive Asset Integrity Monitoring for Industrial Equipment in Oil & Gas and Energy Sector Facilities. International Journal of Scientific Interdisciplinary Research, 5(2), 746–791. https://doi.org/10.63125/3mjcah04

[29]. Mukut Kanti, B. (2025). Lifecycle Carbon Reduction and Net Zero Pathway Modeling Using LEED, EDGE, and SBTI Frameworks in Developing Economy Industrial Buildings. American Journal of Interdisciplinary Studies, 6(3), 306-350. https://doi.org/10.63125/tnneth75

[30]. Nishat, J. (2025). Leveraging Machine Learning to Identify and Address Maternal Health Disparities in Underserved U.S. Communities. American Journal of Data Science and Analytics, 6(12), 125-169. https://doi.org/10.63125/c2t7r719

[31]. Nishat, J., & Mst Kaniz, F. (2022). A Review of the Impact of Covid-19 On Maternal and Child Health Service Utilization in Bangladesh. American Journal of Scholarly Research and Innovation, 1(02), 223–269.

https://doi.org/10.63125/vgv5s742

[32]. Sharif Md Yousuf, B., Md Shahadat, H., Saleh Mohammad, M., Mohammad Shahadat Hossain, S., & Imtiaz, P. (2025). The Green Hydrogen Blueprint: Optimizing America's Clean Energy Future. American Journal of Data Science and Analytics, 6(10), 43-69. https://doi.org/10.63125/y9b1k470

[33]. Tayebi, M., & El Kafhali, S. (2026). Recent progress on financial risk detection in the context of transaction fraud based on machine learning algorithms. Journal of Risk and Financial Management, 19(1), Article 14. https://doi.org/10.3390/jrfm19010014

[34]. Tonay, P., Md Maruf, I., & Al, A. (2024). Artificial Intelligence-Based Decision Support Systems and Managerial Performance. American Journal of Advanced Technology and Engineering Solutions, 4(04), 154-184. https://doi.org/10.63125/wmz8dc67

[35]. Universidad Continental. (2025). Comparative analysis of machine learning models for credit card fraud detection using SMOTE for class imbalance. International Journal of Safety and Security Engineering, 15(5), 913-924. https://iieta.org/journals/ijsse/paper/10.18280/ijsse.150504