The American Journal of Engineering and Technology
ISSN 2689-0984 | Open Access

TYPE Original Research
PAGE NO. 341-347

Adaptive Mechanisms and Game-Theoretic Incentives for Machine Learning-Based Online Payment Fraud Detection: Toward Robust Transactional Integrity

Dr. A. R. Mendes
Global Institute for Financial Technology Studies, Lisbon University

OPEN ACCESS

SUBMITED 01 August 2025
ACCEPTED 15 August 2025
PUBLISHED 31 August 2025
VOLUME Vol.07 Issue 08 2025

CITATION

Dr. A. R. Mendes. (2025). Adaptive Mechanisms and Game-Theoretic Incentives for Machine Learning-Based Online Payment Fraud Detection: Toward Robust Transactional Integrity. The American Journal of Engineering and Technology, 7(8), 341–347. Retrieved from https://theamericanjournals.com/index.php/tajet/article/view/7041

Abstract

Background: The surge in digital commerce and electronic payments has rapidly expanded the attack surface for fraudulent actors, prompting a corresponding acceleration in machine learning approaches to detect and mitigate online payment fraud (Keerthi & Nalini, 2024; Anitha et al., 2025). Simultaneously, scholarship in financial innovation and risk management emphasizes that fraud detection is not merely an engineering problem but a socio-technical challenge involving incentives, market design, and strategic interaction among buyers, sellers, platforms, and adversaries (Vanini et al., 2023; Seera et al., 2024).

Objective: This paper synthesizes empirical, methodological, and theoretical literatures to construct an integrative framework that links machine learning detection systems with game-theoretic incentive mechanisms and governance architectures for transactional integrity. The objective is to articulate design principles for detection systems that are resilient to strategic evasion, that internalize incentive misalignments, and that operate within practicable risk management regimes (Silva et al., 2022; Lucas & Jurgovsky, 2020).

Methods: The study undertakes a systematic synthesis of prior literature on supervised and ensemble machine learning methods for payment fraud detection (Abirami et al., 2018; Singh & Shukla, 2018; Silva et al., 2022), anomaly detection to risk management transition frameworks (Vanini et al., 2023), and incentive-based analyses from game theory applied to transactional integrity (Weiying, 1996; Zhang et al., 2007). The methodology is textual and conceptual: it combines methodological exposition of model families, threat modeling of adversarial behaviors, formalized incentive narratives from game-theoretic literature, and prescriptive governance recommendations.

Results: The integrative framework identifies three core adaptive mechanisms: (1) ensemble and hybrid model architectures to improve detection accuracy and reduce false positives while enabling model diversity against adversarial strategies (Silva et al., 2022; Seera et al., 2024); (2) dynamic risk-scoring systems coupled with economic incentives and reputational mechanisms to align actor behavior in marketplace settings (Zhang et al., 2007; Ma et al., 2005); and (3) governance and operational controls that transform anomaly alerts into enterprise-level risk responses, embedding human-in-the-loop decisioning and escalation pathways (Vanini et al., 2023; Singh, 2025). The framework further details countermeasures to adversarial evasion and discusses trade-offs between privacy, detection performance, and operational cost.

Conclusions: Robust online payment fraud detection requires more than optimized classifiers: it demands an integrated approach combining sophisticated machine learning ensembles, explicit game-theoretic incentive alignment, and institutional risk management strategies. Implementing such systems involves navigating technical, economic, and ethical trade-offs; policy, governance, and continuous monitoring are essential to sustain transactional integrity in evolving digital marketplaces.

Keywords: online payment fraud, machine learning, ensemble methods, game theory, transactional integrity, incentive design, risk management

INTRODUCTION

The rapid proliferation of online payment systems has reshaped commerce, lowering transaction friction and expanding market participation globally. With this digital expansion, the prevalence and sophistication of payment fraud have also grown, creating persistent systemic risks for financial institutions, e-commerce platforms, merchants, and consumers (Keerthi & Nalini, 2024; Anitha et al., 2025). Detection systems that once relied on static rule sets or simple statistical heuristics have been progressively supplanted or augmented by machine learning methods capable of discovering complex patterns in transactional data (Abirami et al., 2018; Lucas & Jurgovsky, 2020). Yet, despite methodological advances, major challenges remain: the adversarial adaptability of fraudsters, severe class imbalance in labeled data, privacy constraints on data sharing, and misaligned incentives across market participants that can undermine the deterrence effects of detection (Vanini et al., 2023; Silva et al., 2022).

The literature on credit card and online payment fraud converges on a set of empirical and theoretical observations. First, supervised learning methods—when well-crafted and trained on representative datasets—can achieve high detection rates, but they are vulnerable to concept drift and targeted evasion strategies (Lucas & Jurgovsky, 2020; Singh & Shukla, 2018). Second, ensemble approaches and hybrid architectures (combining anomaly detection with supervised classifiers) improve robustness by leveraging model diversity and combining complementary detection logics (Silva et al., 2022; Seera et al., 2024). Third, anomaly detection alone cannot substitute for enterprise risk management; turning alerts into operational outcomes requires a calibrated risk-management framework that accounts for false positives, customer experience, and regulatory constraints (Vanini et al., 2023). Finally, game-theoretic analyses of online marketplaces reveal that transactional integrity is as much a function of incentives and reputation mechanisms as it is of technological detection (Weiying, 1996; Zhang et al., 2007; Ma et al., 2005). These observations illuminate the need for an integrated view that situates machine learning-based detection within incentive-aware governance frameworks.

The problem statement at the core of this paper is straightforward but consequential: how can institutions design machine learning-based online payment fraud detection systems that remain effective under strategic adversarial pressure, while aligning economic incentives and preserving operational feasibility? Existing scholarship provides important building blocks but often treats technical detection and incentive analysis in isolation (Silva et al., 2022; Vanini et al., 2023). The literature gap, therefore, is the absence of a fully synthesized framework that operationalizes model architectures, adversarial threat modeling, and game-theoretic incentive mechanisms into coherent, implementable design principles for transactional integrity.

This article proceeds by mapping the methodological landscape of machine learning approaches for online payment fraud detection, elaborating adversarial threat models relevant to these systems, and introducing game-theoretic constructs applicable to transactional ecosystems. Building on that mapping, the paper proposes a set of adaptive mechanisms—ensemble architectures, dynamic risk scoring with incentive alignment, and governance pathways—that together constitute a robust approach to preserving transactional integrity. Throughout, the discussion emphasizes practical trade-offs, technical constraints, and the socio-economic dynamics of marketplaces. The contribution is not an empirical evaluation using new datasets but rather a theoretical and prescriptive synthesis that bridges machine learning practice with economic theory and operational risk management, grounded in contemporary literature (Keerthi & Nalini, 2024; Silva et al., 2022; Vanini et al., 2023; Seera et al., 2024).

METHODOLOGY

The methodological approach is intentionally synthetic and discursive. Rather than presenting new empirical results from original datasets, the paper constructs a conceptual framework derived from critical readings of the referenced literature and from analytical extrapolation of established methods. The methodology is organized into four interrelated components: (1) taxonomy of machine learning methods for fraud detection; (2) adversarial threat modeling and evaluation criteria; (3) incentive and game-theoretic analysis of transactional ecosystems; and (4) governance and operational integration strategies.

Taxonomy of Machine Learning Methods

A comprehensive taxonomy clarifies the methodological building blocks available to practitioners. Supervised learning approaches—logistic regression, decision trees, random forests, gradient boosting machines, and neural networks—have been the backbone of many fraud detection systems due to their ability to leverage labeled historic transactions (Abirami et al., 2018; Singh & Shukla, 2018). Ensemble methods, which combine multiple base learners to reduce variance and bias, have shown empirical success in reducing false negatives and false positives by integrating diverse decision boundaries (Silva et al., 2022). Anomaly detection approaches, ranging from distance-based k-nearest neighbors to density estimation and one-class classifiers, serve to flag transactions that deviate from established behavioral baselines and are valuable when labeled fraud examples are sparse or evolving (Vanini et al., 2023). Hybrid architectures that fuse anomaly detection (for novelty detection) with supervised classification (for pattern-based detection) are increasingly advocated to handle the reality of concept drift and the emergence of new fraud patterns (Seera et al., 2024).

Methodologically, the taxonomy also distinguishes model training paradigms: batch training on historical snapshots, incremental or online learning to accommodate streaming data, and transfer learning when leveraging cross-domain datasets. Each paradigm has trade-offs: batch training can capture comprehensive historic patterns but is slow to react; online learning adapts quickly but may suffer from instability and catastrophic forgetting; transfer learning can provide priors from similar domains but risks negative transfer if distributions diverge (Lucas & Jurgovsky, 2020).

Adversarial Threat Modeling and Evaluation

Adversarial threat modeling formalizes the capabilities, objectives, and constraints of fraudsters. A useful starting taxonomy divides attackers into naive opportunists, organized fraud rings, and advanced persistent fraudsters. Naive opportunists rely on simple heuristics and are typically deterred by basic rule-based systems. Organized rings engage in more sophisticated behaviors—coordinated small-value transactions, synthetic identity creation, and mule networks—requiring detection approaches that analyze network-level patterns and actor relationships (Seera et al., 2024). Advanced persistent fraudsters deliberately probe defenses, conduct adversarial testing, and adaptively craft transactions to evade classifiers, which necessitates robust adversarial-resilient design (Lucas & Jurgovsky, 2020).

Evaluation criteria for models must therefore extend beyond static accuracy metrics to include: robustness to adversarial perturbations, sensitivity to concept drift, false positive and false negative costs in operational terms, latency and computational efficiency, and privacy constraints. Operational evaluation must also simulate attack scenarios where fraudsters have partial knowledge of detection systems, enabling practitioners to stress-test model resiliency (Silva et al., 2022).

Game-Theoretic and Incentive Analysis

Game-theoretic analysis frames the strategic interactions among market participants. The relevant games are typically of incomplete information: buyers and sellers possess private characteristics (quality, type), platforms seek to mediate transactions profitably, and adversaries exploit asymmetries. Classic information economics and mechanism-design insights underpin much of the scholarship on marketplace integrity: incentives and reputational mechanisms can mitigate opportunistic behavior when properly designed (Weiying, 1996; Zhang et al., 2007). Reputation systems, deposit schemes, escrow mechanisms, and dynamic pricing of risk-based fees are all instrumentally relevant in aligning incentives.

The methodological choice here is to represent key interactions as simplified games—e.g., a dynamic signaling game between a seller and a platform—or as repeated games where reputational dynamics accumulate. Equilibrium analysis, albeit stylized, yields prescriptions: when verification costs are low and reputation signals are strong, equilibria with high integrity can be sustained; when verification is costly and reputational mechanisms are noisy, technological detection must substitute for weak incentives (Ma et al., 2005; Zhang et al., 2007).

Governance and Operational Integration

Finally, methodology extends to governance: how to translate alerts into action. This involves designing escalation pathways, human-in-the-loop review processes, feedback loops for model retraining, and regulatory compliance controls. The methodological emphasis is on building governance that is responsive but not brittle: tiered response protocols, differential action thresholds depending on risk scores, and mechanisms for merchant and customer appeal are core components (Vanini et al., 2023; Singh, 2025). Additionally, the paper adopts a socio-technical perspective, recognizing that technologies operate within institutional contexts where incentives, legal frameworks, and organizational capacities shape outcomes.

RESULTS

Given the conceptual and synthetic nature of this research, the "results" are presented as the outputs of the integrative analysis: (1) a set of distilled design principles for machine learning-based fraud detection; (2) a typology of adaptive mechanisms that combine technical and economic instruments; and (3) an articulated set of trade-offs and implementation considerations.

Design Principles for Fraud Detection Systems

From the literature synthesis, several design principles emerge as central to robust detection.

1. Model Diversity and Ensembles: Ensemble techniques, including bagging, boosting, and stacking, offer improved performance against noisy and adversarial inputs by combining heterogeneous learners. Silva et al. (2022) demonstrate that ensemble models reduce variance and improve generalization in payment contexts. Seera et al. (2024) further show that diverse model architectures—combining tree-based methods for interpretability with neural networks for pattern recognition—provide complementary strengths. Therefore, systems should employ ensembles not only for performance gains but also for resilience: an attacker exploiting weaknesses in one model may still be flagged by another.

2. **Hybrid Detection: Supervised + Anomaly Detection:** Supervised classifiers excel when labeled fraud examples exist, but anomalies best capture novel or sparse fraud types (Vanini et al., 2023). By layering anomaly scores with supervised risk predictions—e.g., using anomalies to trigger exploratory reviews or to seed semi-supervised learning—systems maintain coverage over both known and emergent fraud modalities (Seera et al., 2024).

3. Continuous Learning and Concept-Drift Awareness: Fraud patterns evolve, often quickly. Lucas and Jurgovsky (2020) emphasize that static models degrade with drift. Incorporating online learning paradigms or scheduled retraining with rolling windows helps maintain relevance. However, continuous learning must be balanced with stability mechanisms—such as replay buffers, regularization, and human review—to avoid amplifying mistaken signals.

4. Cost-Sensitive and Operational Metrics: Accuracy alone is insufficient. False positives incur customer friction and operational cost; false negatives yield direct financial losses. Designing cost-sensitive loss functions and optimizing thresholds based on economic utility produces systems aligned with enterprise objectives (Singh & Shukla, 2018; Silva et al., 2022).

5. Explainability and Human-in-the-Loop Decisioning:

Given regulatory scrutiny and customer-facing consequences, models should provide interpretable explanations for high-risk scores to facilitate human decision-making and appeals. Tree-based models and attention mechanisms can supply features of explanation; moreover, human experts must be embedded in high-stakes escalations (Vanini et al., 2023).

Adaptive Mechanisms: Integrating Economic Incentives and Technical Controls

The paper identifies three adaptive mechanisms that jointly address detection, incentives, and governance.

Mechanism A — Ensemble and Hybrid Detection Stack: The stack uses multiple classifiers (e.g., gradient-boosted trees, random forests, feedforward neural networks) together with anomaly detection subsystems. Diversity is achieved not only through algorithmic heterogeneity but also via different feature sets (transactional features, device fingerprints, network graphs) and temporal horizons (short-term patterns vs. long-term behavior). This mechanism aims to maximize detection breadth while minimizing single-model failure points (Silva et al., 2022; Seera et al., 2024).

Mechanism B — Dynamic Risk Scores with Economic Levers: Risk scores should feed into economic instruments that influence actor behavior. For sellers, this could mean deposit requirements that scale with risk, escrow durations that vary by score, or commission adjustments that internalize predicted fraud costs. For buyers, dynamic authentication requirements (e.g., step-up verification) reflect transaction risk. These instruments are rooted in game-theoretic analysis: by raising the cost of fraudulent behavior or by increasing the expected detection probability and penalty, the equilibrium incentives shift toward honest behavior (Weiying, 1996; Zhang et al., 2007).

Mechanism C — Governance Pathways and Feedback Loops: Detection outputs must integrate with enterprise governance: tiered response levels, manual review processes, merchant/customer communication channels, and model-feedback loops for labeled data incorporation. The governance mechanism formalizes the path from alert to consequence, ensuring that detection translates into deterrence and remediation (Vanini et al., 2023; Singh, 2025).

Trade-offs and Practical Considerations

Applying these mechanisms entails trade-offs. Ensemble systems are computationally heavier and may increase latency (Silva et al., 2022). Dynamic economic levers risk alienating legitimate users if calibrated harshly, thus requiring careful thresholding and transparent policies (Zhang et al., 2007). Governance pathways must consider legal constraints on data processing and consumer protection rules that vary across jurisdictions (Vanini et al., 2023). Furthermore, privacy-preserving architectures—such as federated learning—introduce complexity when data sharing is necessary for cross-institutional pattern recognition.

DISCUSSION

This section interprets the results, explores theoretical implications, addresses limitations, and outlines avenues for future work. The discussion juxtaposes technical efficacy with economic theory and institutional capacity, emphasizing that sustainable integrity requires systemic alignment.

Interpreting the Adaptive Mechanisms within Market Dynamics

The ensemble and hybrid detection architecture addresses the immediate technical challenge of adversarial evasion by increasing detection heterogeneity, but technical robustness alone does not eradicate fraud. Fraudsters respond to incentives, and so detection must be coupled with mechanisms that alter payoff structures. Game-theoretic analyses inform how to design such mechanisms: in signaling games where sellers send quality signals at a cost, augmenting detection reduces the benefits of masquerading because the expected penalty conditional on detection rises (Weiying, 1996; Zhang et al., 2007). Reputation systems further strengthen equilibria in repeated interactions: when platforms publish credible reputational signals, long-term gains for honest behavior can outweigh short-term fraudulent profits (Ma et al., 2005).

The dynamic risk score is potent because it operationalizes the marginal impact of detection into economic terms. For instance, a high-risk seller facing longer escrow periods or higher deposit requirements confronts a higher opportunity cost for fraud. These instruments are analogous to deposit or bond mechanisms in mechanism design that incentivize truthful revelation and discourage opportunism (Zhang et al., 2007). Importantly, calibration matters: overly punitive measures may deter legitimate participation, while weak measures may be insufficient to deter sophisticated actors.

Adversarial Considerations and Model Robustness

Adversarial machine learning literature cautions that attackers with partial knowledge of models can craft perturbations that evade detection (Lucas & Jurgovsky, 2020). Ensembles mitigate but do not eliminate this vulnerability. Defensive strategies include adversarial training (exposing models to crafted adversarial examples during training), monitoring for distributional shifts, and incorporating anomaly detection that is agnostic to specific label patterns (Silva et al., 2022). Operationally, combining automated detection with human review for borderline or high-impact cases can prevent catastrophic errors and feed high-quality labeled examples back to models.

A crucial nuance is the asymmetric cost of errors: false positives harm customers and merchants through friction and reputational damage; false negatives yield financial loss and regulatory scrutiny. Thus, risk optimization must be context-sensitive: regulators or high-value merchant relationships may require lower false-negative tolerance, while low-value, high-volume environments may prefer aggressive thresholds to contain systemic risk. This tailoring aligns with cost-sensitive approaches in supervised learning (Singh & Shukla, 2018).

Privacy, Data Sharing, and Collective Defense

Effective fraud detection often benefits from cross-platform data sharing and collaborative intelligence. Yet legal and privacy constraints limit the free flow of personally identifiable information (PII). Solutions include privacy-preserving protocols—federated learning, secure multi-party computation, and de-identified feature sharing—that enable collaborative model building while maintaining compliance (Vanini et al., 2023). From the governance perspective, platforms must contractually and technically protect shared data while coordinating on signals that indicate emergent fraud waves. The literature suggests that collective defense increases systemic resilience but demands trust frameworks and governance arrangements that align incentives among participants (Seera et al., 2024).

Limitations of the Current Synthesis

Several limitations deserve explicit recognition. First, the paper is conceptual and does not present novel empirical evaluations; the framework needs operational validation on real-world transactional datasets across diverse markets. Second, the references, while comprehensive in certain respects, do not encompass the entirety of rapidly evolving adversarial ML research; implementers should consult specialized technical literature for specific defensive techniques (Lucas & Jurgovsky, 2020; Silva et al., 2022). Third, the game-theoretic models used are stylized and omit some institutional complexities—such as heterogeneous legal jurisdictions, cross-border enforcement challenges, and varying merchant risk appetites—that influence real-world equilibria (Weiying, 1996; Ma et al., 2005). Finally, operational constraints—computational budgets, latency requirements for real-time transactions, and human resource limitations—constrain the immediate applicability of some recommendations.

Future Research Directions

The synthesis suggests several fruitful avenues for future work. Empirical validation of ensemble/hybrid architectures in production environments, with rigorous A/B testing and causal inference methods, would strengthen the evidence base (Silva et al., 2022; Seera et al., 2024). Cross-institutional studies on the effectiveness of dynamic economic levers—escrow policies, deposits, graduated commissions—would illuminate the behavioral responses of merchants and fraudsters and inform mechanism design (Zhang et al., 2007). Research that integrates adversarial ML techniques with game-theoretic modeling—explicitly modeling attackers that optimize evasion subject to economic constraints—could yield attack-resilient architectures. Lastly, exploration of governance models for privacy-preserving collective defense, including legal templates, technical standards, and trust frameworks, would address a key practical barrier to scalable fraud mitigation (Vanini et al., 2023).

CONCLUSION

This paper argues that managing online payment fraud effectively requires an integrated approach that combines machine learning detection, game-theoretic incentive alignment, and robust governance mechanisms. Ensemble and hybrid architectures offer technical resilience to evolving fraud patterns, while dynamic economic levers and reputational mechanisms reshape the incentives that underpin adversarial behavior. Governance pathways translate detection into operational action, balancing detection sensitivity with customer experience and regulatory compliance. These components are mutually reinforcing: technical detection improves the credibility of reputational signals, economic levers magnify the deterrence effect of detection, and governance ensures accountability and continuous learning.

The security of digital transactions is a moving target. Fraudsters innovate and adapt; detection systems must therefore be adaptive both technically and institutionally. Implementing the proposed framework requires investment in model diversity, data governance, privacy-preserving collaboration, and policy design. While no single architecture guarantees invulnerability, the integrative approach outlined here substantially raises the bar for adversaries and aligns market incentives toward sustained transactional integrity.

REFERENCES

1. M. N. Naga Keerthi and S. Nalini, "Online payment fraud detection using machine learning," Int. J. Creative Research Thoughts (IJCRT), vol. 12, no. 8, pp. a25–a26, Aug. 2024.
2. V. Anitha, Ch. Siri, K. Sai Meghana, M. Joshna, and G. Akanksha, "A survey on online payment fraud detection techniques using machine learning algorithms," Int. J. Res. Appl. Sci. Eng. Technol. (IJRASET), vol. 13, no. 1, pp. 1003–1004, Jan. 2025.
3. S. S. R. Abirami, K. S. Abirami, and S. S. Abirami, "Online payment fraud detection using machine learning," J. Adv. Comput. Sci. Technol., vol. 7, no. 3, pp. 45–50, Mar. 2018.
4. Y. Lucas and J. Jurgovsky, "Credit card fraud detection using machine learning: A survey," arXiv preprint arXiv:2010.06479, Oct. 2020.
5. P. Vanini, S. Rossi, E. Zvizdic, and T. Domenig, "Online payment fraud: From anomaly detection to risk management," Financial Innovation, vol. 9, no. 1, article 66, Mar. 2023.
6. M. Seera, C. P. Lim, A. Kumar, L. Dhamotharan, and K. H. Tan, "An intelligent payment card fraud detection system," Ann. Oper. Res., vol. 334, pp. 445–467, Mar. 2024.
7. J. Silva, R. de Oliveira, and L. O. de Souza, "A machine learning approach for online payment fraud detection using ensemble techniques," Procedia Computer Science, vol. 205, pp. 1350–1357, 2022.
8. S. K. Singh and R. Shukla, "Credit Card Fraud Detection Using Supervised Learning Approach," International Journal of Computer Applications, vol. 180, no. 29, pp. 1–8, 2018.
9. Weiying, Game Theory and Information Economics [M], Shanghai Joint Publishing Shanghai People's Publishing House, 1996.
10. Singh, V., Securing Transactional Integrity: Cybersecurity Practices in Fintech and Core Banking, QTanalytics Publication (Books), 2025, pp. 86–96.
11. Zeng Yong and XU Mao Wei, "Commerce between buyers and sellers in credit mode select Game Theory Analysis," Technology Progress and Policy, 2004.
12. Ma Huimin, Ruo-Bing and Ruo, "Accounting Ruo commerce market operators Reputation Effect Game Analysis," Wuhan University of Technology, 2005.
13. Zhang E., Yang Fei, Wang Ying Luo, "Online trading transaction integrity Incentive Mechanism Design," Journal of Management, 2007.
14. Hongqiong, "C2C trading patterns integrity," Anhui University master's degree thesis, 2009.
15. Xin Zhijie, "Game theory to analyze how to ensure the efficient development of C2C market," Hunan Radio and Television University, 2012.
16. Wang Junyi and Cao Liming, "Based on imperfect information game online shopping trust problem analysis," Computer and Digital Engineering, 2008.
17. Tian Jiuling, "Network Problems on shopping Integrity," Business Economics, 2010.
18. GAO Yan, "Online shopping in the study of social integrity - based on imperfect information dynamic game theory," Huaihai Institute of Technology, 2012.
19. Licheng Chen, "C2C transaction integrity of Game Analysis and Research," Hefei University master's degree thesis, 2009.