INTERNATIONAL JOURNAL OF INTELLIGENT DATA AND MACHIN LEARNING (IJIDML-3087-4262)
eISSN: 3087-4264
Volume. 03, Issue. 04, pp. 01-08, April 2026"

A Systematic Review of Machine Learning Approaches For AI-Driven Fraud Detection in Loyalty Programs

Igor Litovsky
Founder & CTO, Mastermind Loyalty
Toronto, Canada

Article received: 14/01/2026, Article Revised: 14/02/2026, Article Accepted: 13/03/2026, Article Published: 06/04/2026

DOI: https://doi.org/10.55640/ijidml-v03i04-01

ABSTRACT

This article examines machine learning approaches used for fraud detection in loyalty programs, treating loyalty abuse as a distinct analytical problem rather than a simplified extension of payment fraud. The topic is timely because contemporary loyalty ecosystems combine account-based stored value, omnichannel interaction data, partner integrations, and promotion logic, which together generate heterogeneous fraud patterns and unstable labels. The article aims to systematize the prominent model families used in fraud analytics and determine which are most suitable for loyalty-program environments. The study relies on source analysis, comparative review, conceptual synthesis, and analytical generalization. Recent research on fraud analytics, anomaly detection, graph learning, tabular modeling, behavioral biometrics, explainable artificial intelligence, and adaptive risk estimation is examined. The analytical part identifies the strongest methodological trajectories for loyalty fraud detection, including hybrid pipelines, graph-based modeling, and behavior-aware scoring. The findings apply to program operators, fraud analysts, and product teams designing AI-supported decision systems for account protection, redemption control, and abuse prevention.

Keywords: loyalty programs, fraud detection, machine learning, graph neural networks, anomaly detection, tabular learning, behavioral biometrics, explainable AI, risk scoring, promotional abuse

Introduction

Loyalty programs have moved far beyond their earlier marketing function. In current practice, they operate through mobile applications, e-commerce channels, partner interfaces, referral mechanics, tier logic, and redemption workflows that generate a persistent digital value layer around the customer account. Fraud in such environments does not reduce to a single event, such as unauthorized login. It appears through account takeover, fake account farming, coordinated promotional abuse, returns-related manipulation, mileage pooling misuse, suspicious redemption chains, and blended outsider–insider scenarios. Because the underlying data are event-driven, relational, and only partly labeled, loyalty fraud creates a more complex analytical setting than conventional rule-based control can handle.

The article aims to identify, classify, and analytically evaluate machine learning approaches suitable for AI-driven fraud detection in loyalty programs. Three tasks guide the study:

1) to systematize the principal model families used in fraud analytics and to map them to the most typical abuse patterns in loyalty ecosystems;
2) to determine the data and modeling constraints that shape fraud detection quality in loyalty settings, including sparse labels, delayed confirmation, heterogeneous events, and

interaction dependencies across accounts and devices;

3) to formulate selection criteria for practical deployment, with attention to interpretability, latency, relational structure, and the balance between preventive friction and detection depth.

Scientific novelty lies in shifting the review from a generic financial-fraud frame to a loyalty-specific analytical frame. The article does not treat loyalty abuse as a minor variation of payment fraud. It shows that loyalty data combine transactional, behavioral, and relational signals in ways that change model suitability, feature design, and evaluation priorities. The review, therefore, advances a more precise basis for selecting AI methods in programs where redemption, promotion, account identity, and partner-linked value flows intersect.

Materials and Methods

The source base was selected to cover the main methodological layers required for a loyalty-focused review of fraud analytics. K.G. Al-Hashedi and P. Magalingam [1] provide a broad review of financial fraud detection techniques and establish the baseline distinction between classification, clustering, and outlier-oriented approaches. S.O. Arik and T. Pfister [2] develop an interpretable deep-learning architecture for tabular data, which is highly relevant for structured loyalty features derived from transactions, profiles, redemptions, and campaign events. F. Carcillo, Y.-A. Le Borgne, O. Caelen, and G. Bontempi [3] examine the combination of unsupervised and supervised learning in fraud settings characterized by rare labels and shifting patterns. D. Cheng, Y. Zou, S. Xiang, and C. Jiang [4] synthesize graph neural network approaches for fraud detection and clarify why relational modeling becomes decisive when abuse propagates through shared entities. Y. Gorishniy, I. Rubachev, V. Khrulkov, and A. Babenko [5] reassess deep learning for tabular data and show that model superiority on structured features cannot be assumed in advance. G. Pang, C. Shen, L. Cao, and A. van den Hengel [6] review deep anomaly detection and support the analytical treatment of weakly labeled or previously unseen loyalty abuse. M. Papaioannou, F. Pelekoudas-Oikonomou, G. Mantas, E. Serrelis, J. Rodriguez, and M.-A. Fengou [7] analyzes quantitative risk estimation in adaptive authentication, which is relevant for account takeover and suspicious session escalation. I.H. Sarker [8] contributes a compact but functional typology of machine-learning paradigms and deployment logic. G. Vilone and L. Longo [9] systematize explainability concepts and evaluation approaches, supplying a basis for discussing accountable fraud scoring. J. Zhang and Y. Wang [10] review behavioral biometric authentication on smartphones and support the inclusion of device- and interaction-level signals in loyalty fraud analytics.

The study applies comparative analysis, source analysis, conceptual structuring, synthesis, and analytical generalization. These methods are used to classify model families, compare their suitability under loyalty-specific data conditions, and derive applied conclusions for AI-assisted fraud detection in redemption, account, and promotion workflows.

Results

A review of recent machine-learning literature shows that fraud-detection methods cannot be transferred to loyalty programs without adjusting the analytical unit. In card fraud, the unit is often the transaction. In loyalty ecosystems, the relevant unit shifts across account, session, event sequence, promotion cycle, referral chain, device cluster, and redemption pathway. Such variability changes what the model must learn. A supervised classifier trained on isolated events may perform adequately on known account-takeover patterns, yet remain weak on coordinated promo abuse or synthetic-account farming because those behaviors emerge through relationships and repetition rather than a single anomalous transaction [1, 4].

The first large family of approaches remains supervised learning on structured tabular data. This family includes gradient-boosting models, random forests, neural architectures for tabular learning, and related classifiers trained on labeled fraud outcomes. Its practical strength lies in operational simplicity. Loyalty programs already generate variables such as redemption amount, balance change, time since previous login, number of failed attempts, device novelty, promo frequency, refund linkage, and distance between normal and current usage location. Models based on such features are suitable when the program has historical case labels and when the organization needs low-latency scoring at the point of redemption or account recovery. The literature reviewed in [2, 5, 8] suggests, though, that structured-data tasks do not automatically favor deep learning. For many fraud settings, performance depends more on feature quality, temporal framing, and label reliability than on architectural novelty alone. In loyalty programs, this observation has direct force because many harmful events are rare, delayed, or manually reclassified after review.

A second methodological family consists of anomaly-detection approaches. These models become relevant when fraud labels are scarce, noisy, or incomplete. Loyalty abuse frequently exhibits precisely that property. Promotional misuse may be recognized only after a pattern becomes economically visible; suspicious transfers can look legitimate until linked accounts are examined together; returns abuse may surface after lagged reversal logic is analyzed. In such settings, one-class learning, reconstruction-based models, density estimation, and deep anomaly detection provide a way to score departures from expected behavior without requiring a mature fraud label inventory [6]. Their analytical value in loyalty environments is high for early warning, cold-start deployment, and monitoring newly launched promotions. Their limitation is equally explicit: anomaly does not necessarily mean fraud. A new travel season, a major campaign, or a cross-brand partnership can shift normal redemption behavior. For that reason, anomaly detection is more robust when used as a triage layer rather than as the sole decision authority.

A hybrid pipeline, therefore, appears more suitable than single-family models for most loyalty programs. The literature on combining unsupervised and supervised learning shows why this design has remained attractive in fraud settings [3]. Unsupervised or weakly supervised methods can surface suspicious clusters, abnormal sessions, or new patterns of promo interaction. In contrast, downstream supervised models can rank or confirm risk on cases closer to known fraud classes. This architecture fits loyalty operations particularly well. Fake-account creation, referral abuse, or coupon harvesting often begin as weak signals scattered across registration events, browser fingerprints, address fragments, and timing similarities. A hybrid pipeline can first narrow the candidate space, then apply stricter classification to redemptions, points transfers, or voucher issuance. The gain here is not merely computational. It reduces reliance on fully validated fraud labels and enables fraud operations to learn from near-fraud behavior.

The review literature increasingly points to graph-based modeling as the most substantive advance for coordinated fraud. In loyalty ecosystems, abuse frequently spreads through relations among accounts, devices, payment instruments, shipping addresses, IP ranges, household claims, partner bookings, and referral links. This structure is poorly represented in flat records. Graph learning, and graph neural networks in particular, provide a means of encoding connections and propagating risk through neighborhoods rather than treating each action in isolation [4]. Such models are analytically promising for mileage pooling abuse, reward resale chains, mule-account structures, referral rings, and synthetic-account networks. They are instrumental when several events appear benign separately but become suspicious once their shared infrastructure is revealed. The strength of graph approaches lies in structural sensitivity. Their weakness lies in implementation complexity, data integration burden, and the greater difficulty of explaining why a node or edge triggered an intervention. For loyalty operators, graph learning is most defensible where abuse is relational by design and where the data architecture can preserve links among customers, devices, redemptions, and partner endpoints.

Behavior-aware methods form another relevant trajectory. Recent work on adaptive risk estimation and smartphone behavioral biometrics indicates that fraud detection benefits from signals generated during the session itself, not only before it begins [7, 10]. In loyalty settings this insight matters because many high-loss events begin with a technically valid login. Stolen credentials, password reset abuse, or low-friction access via reused passwords may not look suspicious at first glance. The risk becomes clearer when the user behavior diverges from historical interaction patterns: unusual navigation, abrupt transition to redemption, unfamiliar device handling, atypical typing rhythm, or rapid changes in destination details. Session-aware modeling is therefore more suitable than static login control for protecting points, miles, gift cards, and partner rewards. Its practical value is most substantial in mobile-first programs where the account is continuously accessible and where frictionless usage has historically been prioritized.

A further distinction arises between models built for known fraud classes and models built for concept drift. Loyalty programs change constantly. New promotions are launched, rules are revised, partnerships are added, and redemption pathways expand. The resulting fraud surface is unstable. A static classifier trained on last year's abuse cases may miss new forms of campaign gaming. The literature reviewed in [1, 3, 6, 8] supports a design principle in which model maintenance is built into the fraud system rather than treated as an occasional retraining task. For loyalty programs, this translates into rolling baselines, event-window redesign, periodic feature revision, and feedback integration from fraud-review teams. The analytical implication is straightforward: model choice must be tied to the

volatility of the reward logic, not merely to the volume of historical fraud cases.

Interpretability remains a decisive condition because loyalty fraud controls influence customer access, redemption eligibility, and service escalation. A model that blocks a redemption, freezes an account, or challenges a transfer needs a defensible explanation for risk operations and customer support. The explainability literature does not support a simplistic opposition between transparent and accurate systems; it shows instead that explanation quality depends on the type of model, the level of explanation, and the purpose of interpretation [9]. In loyalty programs, explanation serves at least three functions. It supports analyst review, helps calibrate risk policies, and reduces operational conflict between fraud prevention and customer-experience teams. This is where interpretable tabular architectures and explanation-aware model governance become more than technical preferences. They shape whether AI can be used as a stable operational instrument rather than a black-box warning generator.

The literature further indicates that no single metric is sufficient for evaluation. Fraud detection in loyalty programs involves class imbalance, delayed labels, varying loss severity, and differing operational costs. Precision, recall, area under the precision–recall curve, time-to-detection, analyst workload, and false-positive burden all matter, but in different ways across fraud types. Returns abuse and promo abuse impose one cost profile; account takeover with immediate redemption imposes another. Evaluation should therefore be aligned with business impact rather than abstract benchmark performance. Recent work on fraud analytics and tabular modeling supports that conclusion by showing that the framing of dataset construction and deployment strongly influences what model quality actually means [1, 3, 5].

The main analytical synthesis emerging from the review is that loyalty fraud detection is best approached as a layered modeling problem. Supervised tabular models remain effective for well-labeled, repetitive tasks; anomaly detection supports weak-label and cold-start cases; graph methods address coordinated and relational abuse; behavioral and adaptive-risk models strengthen account and session protection; and explainability tools remain necessary for production governance [2–10]. The strongest design direction is therefore hybrid rather than monolithic. Figure 1 summarizes this analytical positioning.

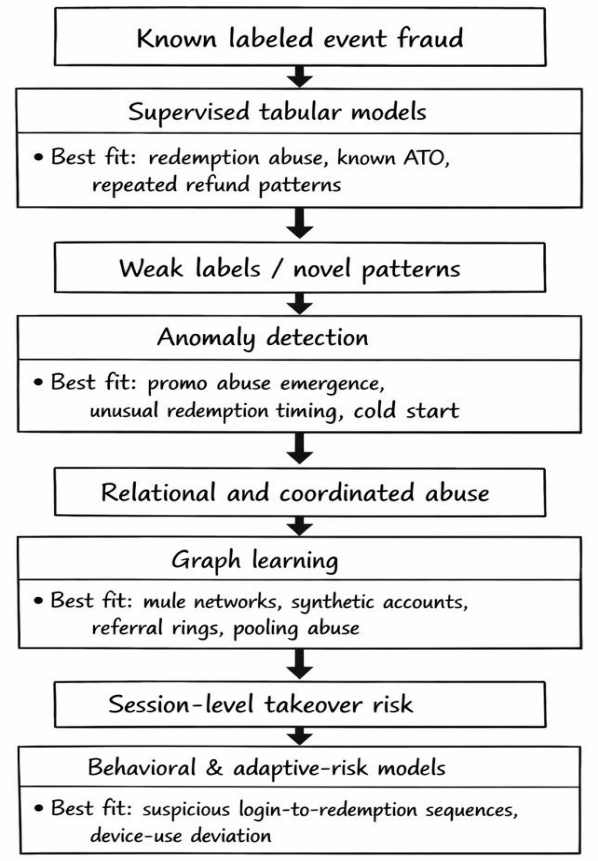

Figure 1. Analytical positioning of machine-learning families for loyalty fraud detection (adapted from [1, 4, 6])

The analytical synthesis presented in Figure 1 shows that fraud detection in loyalty programs cannot be reduced to a single modeling strategy. The reviewed literature indicates that model suitability depends on the structure of the fraud pattern, the availability and quality of labels, and the degree of relational or behavioral complexity embedded in program data. Supervised tabular models remain effective for recurrent and well-documented abuse cases, whereas anomaly detection is more appropriate for weakly labeled or emerging schemes. Graph-based methods offer more substantial analytical value in coordinated multi-account scenarios, while behavioral and adaptive-risk models enhance detection at the session level. Taken together, these findings support a layered detection architecture in which different machine-learning families are aligned with distinct fraud mechanisms rather than applied as interchangeable solutions.

Discussion

The analytical picture developed above suggests that the central question for loyalty programs is not whether AI should replace rule engines. The more precise question concerns where machine learning adds discriminative value beyond rules and where rules remain structurally preferable. Table 1 organizes that distinction by linking common loyalty-fraud scenarios with the model families most likely to perform well under realistic operational conditions [1–10].

Table 1. Preferred machine-learning families by loyalty-fraud scenario [1–10]

Loyalty-fraud scenario	Dominant data structure	Most suitable model family	Main advantage	Main limitation
Account takeover with rapid redemption	Session, device, account history	Supervised tabular + behavioral risk models	Fast inline scoring with strong session sensitivity	Requires good behavioral baselines and reliable escalation logic
Fake account farming	Registration sequences, device reuse, referral ties	Anomaly detection + graph learning	Works under weak labels and reveals coordinated creation patterns	High review burden if anomaly thresholds are too broad
Promotional abuse and multi-account coupon harvesting	Campaign events, identity fragments, shared infrastructure	Hybrid pipeline with anomaly screening and supervised confirmation	Detects both novel and known abuse patterns	Model drift is likely when campaign rules change frequently
Mileage pooling and partner-transfer abuse	Multi-entity relational network	Graph neural networks or graph-based risk propagation	Captures distributed fraud hidden across linked accounts	Demands an integrated graph-ready data architecture
Returns abuse after reward accrual	Event sequence, purchase-return linkage	Supervised tabular + temporal features	Clear use of transaction chronology and reversal lag	Label confirmation may be delayed

Insider-assisted manipulation	Admin actions, override logs, linked customer outcomes	Supervised anomaly scoring on privileged behavior	Targets rare but high-impact misuse	Sparse positive labels reduce direct learnability

The table indicates that the most significant gains from machine learning occur in scenarios where a single event cannot fully capture fraud. Loyalty abuse often unfolds through timing, repetition, linkage, or sequence. That property weakens static rule systems because rules struggle to preserve sensitivity across changing campaigns without generating excessive friction. At the same time, the table shows that model complexity should not be treated as an end in itself. Graph learning is analytically attractive for pooling abuse and account networks, yet it adds value only where relation-rich data are actually available. In smaller programs with narrow fraud taxonomies, interpretable tabular models may remain the better production choice [2, 4, 5].

A second applied question concerns deployment criterion. In loyalty operations, technical quality alone is insufficient. A model has to fit review capacity, latency requirements, customer-experience tolerance, and the type of evidence available for analyst escalation. Table 2 summarizes the main deployment criteria derived from the reviewed sources [2–10].

Table 2. Deployment criteria for AI-driven fraud detection in loyalty programs [2–10]

Criterion	High-priority condition	Better model choice	Why this choice fits
Structured historical case data	Stable fraud labels, repeated patterns	Supervised tabular models	Efficient on account and event features; easier to operationalize
Sparse or noisy fraud labels	New program, weak analyst confirmation	Anomaly detection or a hybrid pipeline	Reduces dependence on mature label inventories
Strong relational dependence	Shared devices, addresses, referrals, partner links	Graph-based models	Preserves network structure that flat tables lose
Mobile-first member behavior	Rich session telemetry, device interaction data	Behavioral risk models	Strengthens detection after a valid login and before redemption
High need for analyst interpretability	Frequent manual review, customer disputes	Interpretable tabular models + explanation layer	Supports defensible decisions and calibration
Fast-changing campaign rules	Frequent promo redesign, seasonal incentives	Hybrid design with rolling retraining	Better resilience to concept drift than static one-shot classifiers

The comparative reading of these criteria leads to a broader conclusion: loyalty fraud detection should be designed as a model portfolio problem, not as a search for a universally superior architecture. That conclusion

follows directly from the diversity of abuse mechanisms. Account takeover, synthetic-account farming, and promo gaming differ not only operationally but analytically: they produce distinct data geometries. The literature on anomaly detection, graph learning, and tabular modeling supports this differentiated view [2–6]. The literature on adaptive risk estimation and behavioral biometrics provides a further correction by showing that the account session itself is a source of predictive evidence, especially when formal authentication has already been completed [7, 10].

The analytical format of the present article has its own limits. It does not test models on a shared loyalty dataset and therefore cannot rank architectures empirically. It does not estimate threshold trade-offs, reviewer capacity, or campaign-specific false-positive costs. It does, however, clarify which methodological choices are theoretically defensible for the loyalty domain and which transfer poorly from generic fraud benchmarks. For practical deployment, this clarification matters. It helps prevent a recurring error in fraud programs: adopting an attractive model family before defining the fraud unit, the label regime, and the operational purpose of the score.

For the professional field surrounding loyalty operations, the applied consequence is direct. AI should be embedded to distinguish suspicious behavior from normal loyalty engagement without causing indiscriminate friction. That favors layered scoring, event-to-account linkage, graph-aware monitoring for coordinated abuse, and explanation-ready decisioning for analyst teams. In other words, the most suitable AI design for loyalty programs is one that respects the data shape of the fraud mechanism being targeted.

Conclusion

The first task of the article was to systematize the principal machine-learning families used in fraud analytics and relate them to loyalty abuse patterns. This task was fulfilled by distinguishing supervised tabular models, anomaly-detection methods, graph-based approaches, behavioral risk models, and hybrid pipelines, and by showing that each family corresponds to different fraud geometries within loyalty programs.

The second task was to identify the data and modeling constraints that determine the quality of fraud detection in loyalty ecosystems. The analysis showed that sparse labels, delayed confirmation, relational dependence, session-level deviation, and rapid rule changes are the main constraints that shape model suitability. Loyalty fraud, therefore, cannot be handled adequately through flat event classification alone.

The third task was to formulate model-selection criteria for deployment. The article established that interpretable tabular models remain effective for stable, known fraud classes; anomaly detection is better suited to weak-label and cold-start settings; graph learning is preferable for coordinated multi-entity abuse; behavioral modeling strengthens protection against session-level takeover; hybrid pipelines offer the strongest overall fit where loyalty fraud types coexist.

The review supports a clear conclusion: AI-driven fraud detection in loyalty programs is most effective when the model choice aligns with the structure of the abuse mechanism, the maturity of the labels, and the operational use of the score. A layered analytical design provides a stronger foundation than any single-model solution.

References

1. Al-Hashedi, K. G., & Magalingam, P. (2021). Financial fraud detection applying data mining techniques: A comprehensive review from 2009 to 2019. *Computer Science Review, 40*, 100402. https://doi.org/10.1016/j.cosrev.2021.100402
2. Arik, S. Ö., & Pfister, T. (2021). TabNet: Attentive interpretable tabular learning. *Proceedings of the AAAI Conference on Artificial Intelligence, 35*(8), 6679–6687. https://doi.org/10.1609/aaai.v35i8.16826
3. Carcillo, F., Le Borgne, Y.-A., Caelen, O., Kessaci, Y., Oblé, F., & Bontempi, G. (2019). Combining unsupervised and supervised learning in credit card fraud detection. *Information Sciences*. Advance online publication. https://doi.org/10.1016/j.ins.2019.05.042
4. Cheng, D., Zou, Y., Xiang, S., & Jiang, C. (2025). Graph neural networks for financial fraud detection: A review. *Frontiers of Computer Science, 19*(9), 199609.
5. Gorishniy, Y., Rubachev, I., Khrulkov, V., & Babenko, A. (2021). Revisiting deep learning models for tabular data. *Advances in Neural Information Processing Systems, 34*, 18932–18943.
6. Pang, G., Shen, C., Cao, L., & Van Den Hengel, A. (2021). Deep learning for anomaly detection: A review. *ACM Computing Surveys, 54*(2), 1–38.
7. Papaioannou, M., Pelekoudas-Oikonomou, F., Mantas, G., Serrelis, E., Rodriguez, J., & Fengou, M.-A. (2023). A survey on quantitative risk

estimation approaches for secure and usable user authentication on smartphones. *Sensors, 23*(6), 2979. https://doi.org/10.3390/s23062979

8. Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science, 2*(3), 160. https://doi.org/10.1007/s42979-021-00592-x

9. Vilone, G., & Longo, L. (2021). Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion, 76*, 89–106. https://doi.org/10.1016/j.inffus.2021.05.009

10. Zhang, J., & Wang, Y. (2024). A survey of behavioral biometric authentication on smartphones. In *Proceedings of the 2023 4th International Conference on Machine Learning and Computer Application (ICMLCA '23)* (pp. 722–729). Association for Computing Machinery. https://doi.org/10.1145/3650215.3650342