Cureus Journal of Computer Science
https://doi.org/10.7759/s44389-026-00149-0

Review Article

Machine Learning for Digital Payment Fraud Detection in Financial Institutions: A Review of Methods, Explainability, and Performance

Moshood Sorinola [1] , Kehinde Hamed [2,]

1. *Finance and Business Analytics, Georgia State University, Atlanta, USA*
2. *Financial Crime and Compliance Management, Utica University, Utica, USA*

Received: June 17, 2026 | Review began: June 21, 2026 | Review ended: June 30, 2026 | Published: July 03, 2026

Abstract

Digital payment fraud has grown alongside real-time and account-to-account payment rails, where fraud types pose distinct detection problems. This narrative review examined peer-reviewed studies on machine learning for digital payment fraud detection in financial institutions, drawing on government and regulator publications only to establish fraud scale and policy context. It assessed detection methods, fraud typologies, reported performance, class-imbalance handling, explainability, and regulatory relevance across 47 sources. Most measurable performance evidence came from credit card studies, where supervised and ensemble models reported strong results that proved sensitive to the dataset, the metric emphasized, and the way validation was performed. For authorized push payment fraud, a high-loss authorized-transfer category that now carries direct reimbursement liability for payment service providers, no peer-reviewed detection-performance study was found; its only quantified figure came from a non-peer-reviewed industry pilot. Account takeover and synthetic identity fraud were similarly underserved. Recurring weaknesses ran deeper than any single method: class imbalance, data-leakage risk, absent temporal validation, concept drift, and accuracy-heavy metrics that need not reflect deployment cost. Explainability was weaker still. Many studies treated predictive accuracy as proof that an explanation was sound, while the resampling used to manage rare fraud can itself distort post hoc tools such as Local Interpretable Model-agnostic Explanations and SHapley Additive exPlanations. Machine learning fraud detection is mature for credit card fraud but not yet validated for authorized-transfer and scam-based fraud. Detection research must be built on institution-held authorized-transfer data, leakage-free protocols, cost-sensitive metrics, and direct tests of explanation fidelity.

Categories: AI applications, Explainable AI, Machine Learning (ML)

Keywords: machine learning, fraud detection, authorized push payment fraud, digital payments, explainable ai, shap, class imbalance, financial institution

Introduction And Background

Digital payment fraud imposes annual losses that run into the tens of billions across major economies. United States consumers reported more than 12.5 billion dollars lost to fraud in 2024, a 25 percent increase over the prior year [1], while criminals in the United Kingdom stole 1.17 billion pounds through authorized and unauthorized fraud during the same year [2]. Global card fraud losses alone reached 32.34 billion dollars in 2021 [3]. As transaction volumes migrate to real-time and account-to-account rails, the speed and irrevocability of these payments, where funds cannot be reversed once deposited, have made them attractive targets for fraudsters [4].

How to cite this article:
Sorinola M, Hamed K (July 03, 2026) Machine Learning for Digital Payment Fraud Detection in Financial Institutions: A Review of Methods, Explainability, and Performance. Cureus J Comput Sci 3 : es44389-026-00149-0. DOI
https://doi.org/10.7759/s44389-026-00149-0

Two categories of payment fraud impose fundamentally different detection problems. Unauthorized fraud, where a criminal transacts without the account holder's involvement, can be detected and stopped by a financial institution monitoring for irregular activity, and the prevention figures reflect this: banks intercept a substantial share of unauthorized fraud before loss occurs. Authorized push payment fraud cannot be intercepted the same way: the victim is deceived into initiating the payment, and a financial institution will ordinarily complete the transfer because the payment instruction carries valid authorization [4]. A single deception underlies a family of related typologies, including account takeover, where credentials are compromised, and synthetic identity fraud, where fabricated identities pass onboarding checks, each of which presents a detection surface different from the anomalous-transaction pattern that credit card models are built to recognize. This authorized-payment distinction has also acquired direct financial consequences. The United Kingdom mandatory reimbursement scheme, in force from October 2024, shifts the cost of authorized push payment fraud onto the payment service provider [5], converting detection of this fraud type from a customer-service concern into a balance-sheet liability.

The machine learning response to payment fraud has developed along several method families, each resting on a distinct learning assumption. Supervised classifiers trained on labeled transactions remain the most widely used, spanning logistic regression and support vector machines through tree ensembles such as random forest, AdaBoost, and the gradient-boosted variants XGBoost and CatBoost [6-9]. Hybrid pipelines pair these classifiers with information-theoretic or wrapper-based feature selection to sharpen separability on skewed data [10]. Deep learning has extended the field toward representations that model the temporal structure of spending: long short-term memory networks and attention mechanisms recast fraud as a sequence-classification problem rather than independent transaction scoring [11-13], and generative adversarial networks synthesise minority-class examples that represent the fraud distribution more faithfully than simple duplication [14]. Unsupervised and graph-based methods relax the labeled-data requirement, with isolation forests flagging anomalies through random partitioning [15] and graph neural networks modeling the relational structure of transaction networks [16]. Every method family must also contend with class imbalance, for which the synthetic minority over-sampling technique and its descendants are the standard correction [17], and with the post-hoc explainability tooling (chiefly SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME)) that regulated deployment increasingly demands [18-19]. Systematic reviews spanning 50 to more than 120 studies have repeatedly surveyed this literature, converging on two observations: deep learning remains comparatively underexplored relative to classical methods, and class-imbalance handling is unresolved [6,20-23].

Most of this work evaluates detection on credit card transaction benchmarks, where labeled data is abundant and the fraudulent signal is statistically separable. The reported performance of these credit-card methods has not been tested against authorized push payment fraud, account takeover, or synthetic identity fraud, and the explainability tooling they rely on has not been shown to satisfy the validation that financial regulation requires.

This review examines machine learning methods for digital payment fraud detection in financial institutions, the performance reported across fraud types and datasets, and the explainability and evaluation practices that govern deployment. We organize the evidence by fraud type and method family, consolidate reported performance into a comparison across studies, and assess where the concentration of research effort aligns with, and diverges from, the distribution of financial loss. The review's contribution is to identify authorized-transfer fraud, now a direct reimbursement liability for payment service providers, as a category the peer-reviewed literature has left without a detection benchmark, a validated method, or a reported performance figure.

Review

Methods

This narrative review used a structured but non-systematic search to assemble peer-reviewed evidence on machine learning for digital payment fraud detection. Six databases were consulted, IEEE Xplore, ScienceDirect, SpringerLink, MDPI, PLOS, and Google Scholar, together with the reference lists of retrieved reviews and targeted citation tracking of individual studies identified by name. Search terms combined a fraud-domain group, a method group, and an evaluation

group in the Boolean form shown below; the expression was adapted to the field syntax of each database (for example, title and abstract restriction in IEEE Xplore and ScienceDirect and unrestricted full-text matching in Google Scholar). The search covered work published between 2002 and 2026, with foundational method papers predating that window retained for their definitional role. The search and selection process was not registered as a protocol, did not apply dual independent screening, and did not record per-database retrieval yields. Figure *1* accordingly sets out the search strategy and the selection criteria applied rather than a quantitative screening cascade.

Search string (core Boolean expression, adapted per database): ("credit card fraud" OR "authorized push payment fraud" OR "account takeover" OR "synthetic identity fraud" OR "payment fraud") AND ("machine learning" OR "deep learning" OR "neural network" OR "ensemble" OR "anomaly detection" OR "explainable AI" OR "SHAP" OR "LIME") AND ("precision" OR "recall" OR "AUC" OR "performance").

Inclusion was restricted to peer-reviewed journal articles and conference proceedings for any quantified detection-performance claim entering the comparative analysis. Government and financial-regulator publications were admitted for fraud-scale figures, regulatory context, and operational examples but were excluded from the comparative performance table, which draws only on peer-reviewed sources. Studies were retained when they reported a quantified detection outcome, defined a fraud typology, or established a method or evaluation practice relevant to digital payment fraud in financial institutions.

Preprints without subsequent peer-reviewed publication were excluded from all performance and method claims entering the comparative analysis, as were vendor white papers, marketing materials, and articles in non-indexed venues. Where a study existed in both preprint and peer-reviewed form, the peer-reviewed version was used. Metadata for every retained source was verified against the published record, and reported performance figures were extracted directly from the source text rather than from secondary citation. The structured search retained 47 peer-reviewed and foundational sources spanning the fraud landscape, method and performance, explainability and regulatory evidence, and emerging approaches, sequence transformers, federated learning, and large language models, all admitted under the same inclusion criteria. Figure *1* summarises the search strategy and the selection criteria applied.

How to cite this article:
Sorinola M, Hamed K (July 03, 2026) Machine Learning for Digital Payment Fraud Detection in Financial Institutions: A Review of Methods, Explainability, and Performance. Cureus J Comput Sci 3 : es44389-026-00149-0. DOI https://doi.org/10.7759/s44389-026-00149-0

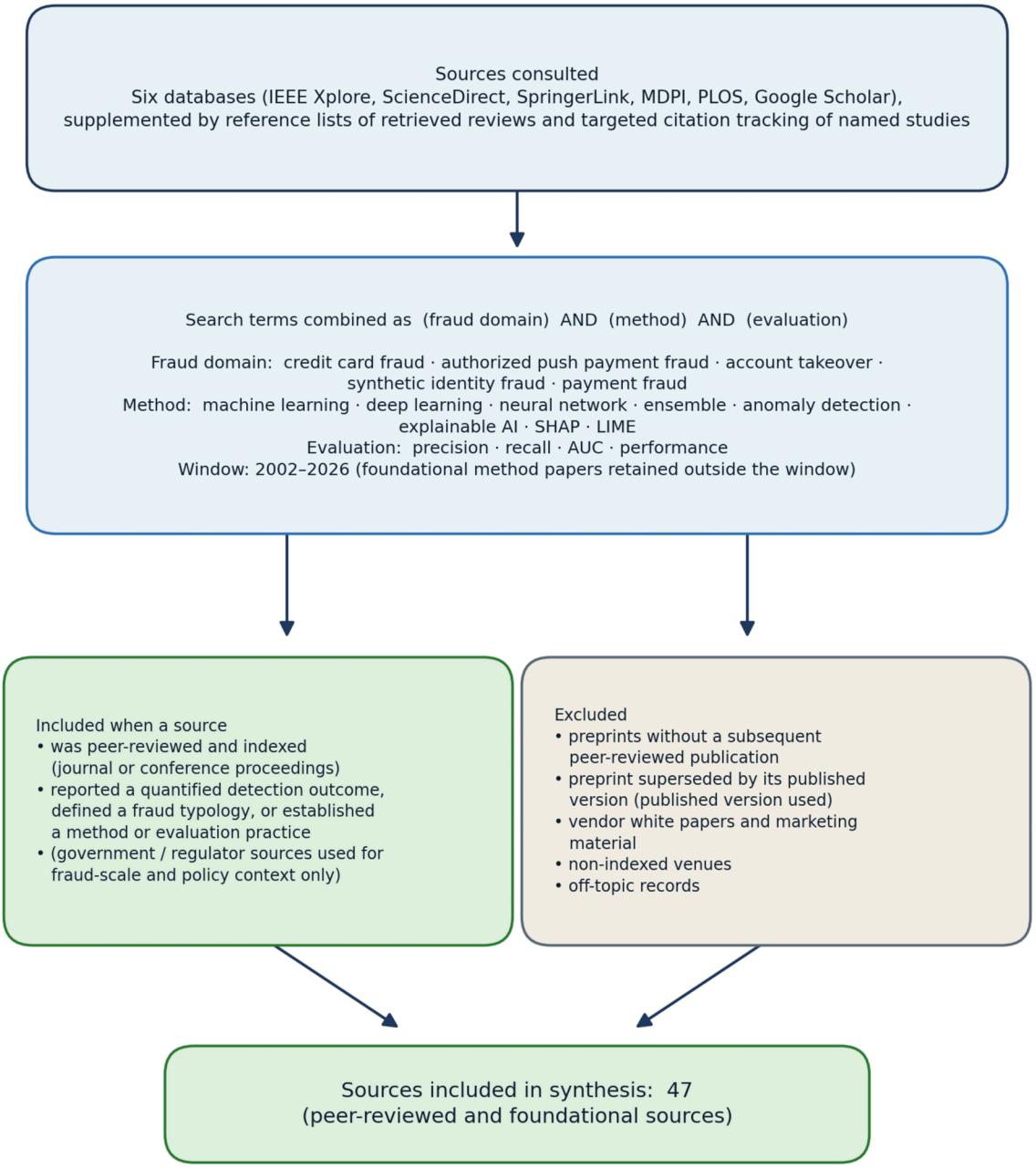

FIGURE 1: Search strategy and source-selection criteria for the narrative review

AUC, Area Under the Curve; LIME, Local Interpretable Model-agnostic Explanations; SHAP, SHapley Additive exPlanations

Results

Digital Payment Fraud Landscape

How to cite this article:
Sorinola M, Hamed K (July 03, 2026) Machine Learning for Digital Payment Fraud Detection in Financial Institutions: A Review of Methods, Explainability, and Performance. Cureus J Comput Sci 3 : es44389-026-00149-0. DOI https://doi.org/10.7759/s44389-026-00149-0

The fraud types responsible for the largest losses are not those most amenable to detection, a mismatch rooted in how each is carried out. Contemporary digital payment fraud spans several distinct typologies, each defined by how the fraudster obtains value. Account takeover involves unauthorized access to a legitimate user account through stolen credentials or social engineering, enabling fraudulent transactions under the guise of the genuine account holder [24]. Authorized push payment fraud operates differently: the victim is deceived into initiating a payment to a fraudster, typically through impersonation or urgency-based manipulation [24], and the victim authorizes the transfer of funds from their own account believing the transaction to be legitimate [5]. Synthetic identity fraud constructs fictitious identities by combining real and fabricated personal data, allowing fraudsters to open accounts and obtain credit undetected [24]. Money laundering recruits individuals to move illicit funds across accounts through instant payments, SEPA transfers, and prepaid card top-ups, often signaled by device or SIM changes and recurring beneficiary patterns [24]. Within authorized push payment frauds, a widely adopted taxonomy distinguishes eight scam types grouped under Malicious Payee and Malicious Redirection headings [5]. A complementary framework classifies payment scams along two dimensions, authorized versus unauthorized and push versus pull, yielding four categories [4].

Across jurisdictions, the financial scale of these typologies has grown. In the United Kingdom, criminals stole 1.17 billion pounds through authorized and unauthorized fraud during 2024 [2], of which authorized push payment fraud accounted for 450.7 million pounds across fewer than 186,000 cases [2]. That 2024 total followed 485.2 million pounds lost to authorized push payment fraud across 207,372 incidents in 2022 [5]. United States consumers reported losing more than 12.5 billion dollars to fraud in 2024, a 25 percent increase over the prior year [1], and lost more money through bank transfers and cryptocurrency than through all other payment methods combined [1]. Disputes at the three largest banks participating in Zelle exceeded 206 million dollars during 2023, with scam victims bearing more than 80 percent of those losses [4]. Global card fraud losses reached 32.34 billion dollars in 2021 [3], and credit card fraud remains the most extensively researched external fraud type, with projected annual losses exceeding 40 billion dollars by 2027 [24].

Authorized transactions present a distinct detection problem. Unlike an unauthorized payment, which a financial institution may detect and stop, an authorized payment initiated by the victim toward a fraudster will most likely be executed because the payment instruction is genuine [4]. Social engineering coaches victims to bypass anti-fraud checks, rendering even sophisticated controls ineffective [5]. Speed and irrevocability compound the difficulty, since funds in faster payment systems cannot be reversed once deposited in the recipient account, a property that makes these rails attractive to fraudsters [4]. Banks prevented 1.45 billion pounds of unauthorized fraud during 2024 through advanced security systems [2], while authorized push payment losses persisted, with 70 percent of cases originating online and 16 percent through telecommunications networks [2].

Regulatory frameworks address reimbursement for authorized transfers. Authorized payments fall outside the rules governing reimbursement for unauthorized transfers, leaving victims of authorized-transfer fraud in a legal lacuna [5]. United Kingdom legislation introduced a mandatory reimbursement scheme for certain authorized push payment fraud, in force from October 2024 [5], following the 2023 Supreme Court decision in Philipp v. Barclays, which held that a bank's duty of care does not apply where the customer is the victim of authorized push payment fraud because the payment instruction itself is not in dispute [5]. A Payment Systems Regulator policy statement codified this reimbursement requirement for payments sent through Faster Payments [25]. The European Payment Services Directive 2 introduced Strong Customer Authentication for electronic transactions [26]. A non-peer-reviewed industry pilot by Featurespace with Pay.UK, discussed in a Federal Reserve Bank of Kansas City briefing, reported identifying 56 percent of authorized push payment scam transactions, with £180 million in detected fraud and a 5:1 false-positive ratio [4].

Machine Learning Methods and Reported Performance

Most of the quantified performance evidence comes from credit card datasets, and a method's reported scores can change substantially from one dataset to another. Machine learning research applied to payment fraud detection spans supervised, unsupervised, and hybrid approaches across reviews covering between 50 and more than 120 studies [6,20-22]. A systematic review by Cherif et al. [23] of 40 studies published between 2015 and 2021 categorized the field by class imbalance handling, feature engineering, and modeling technique, and concluded that deep learning remained

comparatively underexplored relative to classical methods, identifying it as a direction requiring further work. West and Bhattacharya [6] noted that traditional manual detection methods are time-consuming, expensive, inaccurate, and impractical in the era of big data, which has driven financial institutions toward automated computational approaches. Neural networks and support vector machines have demonstrated particular capability for fraud detection due to their ability to learn and adapt to new fraud patterns, though reported performance varies across fraud types and datasets [6].

Supervised classification on credit card benchmarks dominates the reported performance literature. Mienye and Sun [20] developed a hybrid feature-selection method combining information gain with a genetic algorithm wrapper, achieving a sensitivity of 0.997, a specificity of 0.994, and an area under the curve of 0.990 on the European cardholders dataset. The same method achieved a sensitivity of 0.904, a specificity of 0.945, and an area under the curve of 0.910 on the German credit dataset [20], and a sensitivity of 0.945, a specificity of 0.961, and an area under the curve of 0.940 on the Taiwan credit card dataset [20]. Sensitivity fell from 0.997 to 0.904 across datasets while the method itself remained unchanged.

Tayebi and El Kafhali [27] reported similar instability within a single study: their Bayesian-optimized XGBoost achieved an area under the curve of 0.9879 on one credit card dataset using synthetic oversampling, but on a second dataset the best configuration used random undersampling and achieved only 0.9088, an eight-point swing in which even the optimal resampling strategy differed between datasets.

Du et al. [28] compared an autoencoder-XGBoost architecture against baseline classifiers, reporting an overall accuracy of 0.9993, a true positive rate of 0.7839, a true negative rate of 0.9997, and a Matthews correlation coefficient of 0.8845. Within the same comparison, baseline XGBoost achieved an accuracy of 0.9696, a true positive rate of 0.8529, and a Matthews correlation coefficient of 0.8100 [28]; K-nearest neighbors achieved a true positive rate of 0.8835 and a Matthews correlation coefficient of 0.5903 [28]; and random forest achieved a true positive rate of 0.8025 and a Matthews correlation coefficient of 0.6902 [28]. Among the four, the autoencoder-XGBoost method achieved the highest Matthews correlation coefficient yet the lowest true positive rate, catching fewer fraudulent transactions than K-nearest neighbors despite a stronger aggregate score.

Albalawi and Dardouri [7] reported that XGBoost achieved a precision of 91.67 percent, a recall of 95.00 percent, and an F1-score of 93.30, outperforming classical baselines, compared with a deep learning model with focal loss that achieved a precision of 72.97 percent, a recall of 87.10 percent, and an F1-score of 0.7941 [7]. Logistic regression in the same comparison achieved the highest recall at 90.32 percent but a precision of only 15.73 percent [7], meaning that roughly five out of every six flagged transactions were false positives. Randhawa et al. [8] compared 12 algorithms and found that AdaBoost combined with majority voting achieved higher detection accuracy and greater robustness to noise than individual classifiers.

Beyond the credit card datasets that dominate this literature, a smaller body of work reports performance on other digital payment channels. Hajek et al. [29] evaluated an XGBoost-based framework on the PaySim mobile payment dataset of more than six million transactions, comparing XGBoost against k-nearest neighbors, support vector machines, and random forest. They designed the framework to weigh the financial consequences of detection rather than classification accuracy alone. Almazroi and Ayub [30] addressed online payment fraud with a ResNeXt-embedded gated recurrent unit model evaluated on the PaySim and IEEE-CIS datasets, applying the Synthetic Minority Over-sampling Technique to counter class imbalance. Alarfaj et al. [31] applied a convolutional neural network to the European cardholders dataset and reported, under cross-validation, an accuracy of 0.9986, a precision of 0.9971, a recall of 1.0, and an F1-score of 0.9986 - figures so close to perfect that they raise the question of whether the evaluation protocol, rather than the architecture, produced them.

A separate line of work evaluates methods using cost rather than classification accuracy. Almhaithawi et al. [9] measured a savings metric that weights each transaction by its monetary value and found that CatBoost alone achieved savings of 0.7158, increasing to 0.971 when synthetic oversampling was applied, and to 0.9762 when oversampling was paired with a Bayes minimum-risk cost wrapper, while XGBoost reached its best savings of 0.757 under the cost wrapper without

oversampling. The ranking of methods under the savings metric did not match their ranking under accuracy, which is the central point: a model that appears competitive under aggregate accuracy can perform substantially worse once the financial weight of each missed or falsely flagged transaction is considered.

Baisholan et al. [32] built an interpretable ensemble combining random forest and XGBoost, deliberately retaining the natural class imbalance rather than applying oversampling, and reported a recall of 95 percent and an area under the precision-recall curve of 97 percent, arguing that precision-recall metrics are more reliable than accuracy on imbalanced data because accuracy can be misleading when fraud is rare. Table *1* consolidates these and the remaining metric-bearing studies, listing each method alongside its evaluation dataset, reported performance, and fraud type addressed.

How to cite this article:
Sorinola M, Hamed K (July 03, 2026) Machine Learning for Digital Payment Fraud Detection in Financial Institutions: A Review of Methods, Explainability, and Performance. Cureus J Comput Sci 3 : es44389-026-00149-0. DOI https://doi.org/10.7759/s44389-026-00149-0

Method	Study	Dataset	Key reported metrics	Fraud type
IG-GAW hybrid (feature selection + ELM)	Mienye and Sun [20]	European cardholders	Sensitivity 0.997, Specificity 0.994, AUC 0.990	Credit card
IG-GAW hybrid	Mienye and Sun [20]	German credit card	Sensitivity 0.904, Specificity 0.945, AUC 0.910	Credit card
IG-GAW hybrid	Mienye and Sun [20]	Taiwan credit card	Sensitivity 0.945, Specificity 0.961, AUC 0.940	Credit card
AE-XGB-SMOTE-CGAN	Du et al. [28]	European cardholders	ACC 0.9993, TPR 0.7839, TNR 0.9997, MCC 0.8845	Credit card
XGBoost (baseline)	Du et al. [28]	European cardholders	ACC 0.9696, TPR 0.8529, MCC 0.8100	Credit card
K-nearest neighbors	Du et al. [28]	European cardholders	ACC 0.9691, TPR 0.8835, MCC 0.5903	Credit card
Random forest	Du et al. [28]	European cardholders	ACC 0.9583, TPR 0.8025, MCC 0.6902	Credit card
XGBoost (with SMOTE)	Albalawi and Dardouri [7]	Real-world credit card	Precision 91.67%, Recall 95.00%, F1 93.30	Credit card
Deep learning (focal loss)	Albalawi and Dardouri [7]	Real-world credit card	Precision 72.97%, Recall 87.10%, F1 0.7941	Credit card
Logistic regression	Albalawi and Dardouri [7]	Real-world credit card	Precision 15.73%, Recall 90.32%	Credit card
AdaBoost + majority voting	Randhawa et al. [8]	European + real-world	Higher accuracy and noise resistance than individual classifiers	Credit card
LSTM + attention + UMAP	Benchaji et al. [11]	Credit card transactions	Sequential modeling with feature selection	Credit card
Graph neural network	Motie and Raahemi [16]	Financial transaction graphs	Relational structure modeling; competitive with classical ML	Financial (multiple)
CNN (deep learning)	Alarfaj et al. [31]	European cardholders	ACC 0.9986, Precision 0.9971, Recall 1.0, F1 0.9986 (cross-validation)	Credit card

How to cite this article:
Sorinola M, Hamed K (July 03, 2026) Machine Learning for Digital Payment Fraud Detection in Financial Institutions: A Review of Methods, Explainability, and Performance. Cureus J Comput Sci 3 : es44389-026-00149-0. DOI https://doi.org/10.7759/s44389-026-00149-0

XGBoost framework (cost-sensitive)	Hajek et al. [29]	PaySim mobile payment	XGBoost vs kNN/SVM/RF; financial-cost weighted	Mobile payment
ResNeXt-GRU (RXT)	Almazroi and Ayub [30]	PaySim + IEEE-CIS	Deep hybrid with SMOTE; real-time processing	Online payment
XGBoost + Bayesian opt.	Tayebi and El Kafhali [27]	Credit card (Data 1)	AUC 0.9879 with SMOTE (ACC 0.9996, Precision 0.9406)	Credit card
XGBoost + Bayesian opt.	Tayebi and El Kafhali [27]	Credit card (Data 2)	AUC 0.9088 with under-sampling (ACC 0.8325)	Credit card
CatBoost (cost-sensitive)	Almhaithawi et al. [9]	Real-world credit card	Savings 0.7158 alone, 0.9762 with SMOTE+BMR	Credit card
RF + XGBoost (interpretable)	Baisholan et al. [32]	European cardholders	Recall 95%, AUC-PR 97%; no oversampling, SHAP	Credit card
Isolation forest	Liu et al. [15]	General anomaly detection	Unsupervised; no labeled data required	General
Transformer–BDI (self-attention + knowledge distillation)	Dubey et al. [33]	IEEE-CIS + European + PaySim	Improved AUC and interpretability vs classical/deep baselines; real-time capable	Multiple
Instruction-tuned LLM (FinFraud-LLM)	Chen et al. [34]	IEEE-CIS	ACC 97.51% but Recall 35.53% (below XGBoost recall 59.21%)	Financial (multiple)
Federated transfer learning (FED-SPFD)	Chen et al. [35]	Real-world (Kaggle), distributed	Improved recall under heterogeneous cross-institution data; privacy-preserving	Credit card

TABLE 1: Selected machine learning methods and performance evidence relevant to digital payment fraud detection.

ACC, Accuracy; AE-XGB-SMOTE-CGAN, Autoencoder with XGBoost based on Synthetic Minority Over-Sampling Technique and Conditional Generative Adversarial Network; AUC, Area Under the Curve; BMR, Bayes Minimum Risk; CatBoost, Categorical Boosting; CNN, Convolutional Neural Network; ELM, Extreme Learning Machine; GRU, Gated Recurrent Unit; KNN, K-Nearest Neighbors; LLM, Large Language Model; LSTM, Long Short-Term Memory; MCC, Matthews Correlation Coefficient; RF, Random Forest; SHAP, SHapley Additive exPlanations; SVM, Support Vector Machine; TNR, True Negative Rate; TPR, True Positive Rate; UMAP, Uniform Manifold Approximation and Projection; XGBoost, Extreme Gradient Boosting

Deep learning approaches that model the temporal structure of transactions have addressed limitations of transaction-level classification. Sequence matters. Benchaji et al. [11] used long short-term memory networks to incorporate transaction sequences and capture the sequential nature of transactional data, allowing the classifier to identify the most predictive transactions within an input sequence. That architecture combined uniform manifold approximation and projection for feature selection, long short-term memory for sequence modeling, and an attention mechanism to weight informative transactions [11]. Jurgovsky et al. [12] treated fraud detection as a sequence classification problem rather than independent transaction scoring, improving detection by evaluating each transaction in the context of preceding cardholder behavior. Esenogho et al. [13] built a neural network ensemble using long short-term memory as the base learner within an adaptive boosting implementation. Fiore et al. [14] applied generative adversarial networks to generate synthetic fraudulent transactions that better represent minority class distributions than traditional oversampling.

Unsupervised and graph-based methods extend detection beyond labeled supervised classification. Liu et al. [15] introduced Isolation Forest, which identifies anomalies by random partitioning, isolating anomalous observations in fewer partitions than normal points without requiring labeled fraud data. Motie and Raahemi [16] found that graph neural networks capture the interconnected relationships within financial transaction networks, offering an advantage over traditional methods for modeling relational structure, yet their performance, evaluated using accuracy, precision, recall, F1-score, and area under the curve, does not uniformly exceed that of traditional machine learning and deep learning methods, which may still perform competitively when data is not graph-structured or when the graph is relatively small [16].

Class imbalance is the cross-cutting constraint for every method. Fraud is rare. Most fraud detection datasets exhibit extreme imbalance, with fraudulent transactions typically constituting less than 1 percent of all transactions [24]. Within the European cardholders' benchmark, only 0.172 percent of 284,807 transactions are fraudulent [10,28], a skew that prevents conventional algorithms from generalizing minority class behavior [36]. A synthetic minority over-sampling technique introduced by Chawla et al. [17] addresses this by interpolating between existing minority samples and their nearest neighbors rather than duplicating them.

Feature representation and deployment conditions shape performance beyond algorithm choice. Bahnsen et al. [37] found that aggregated transaction features derived from cardholder spending history, such as average expenditure per period and deviation from typical patterns, improve detection beyond raw transaction attributes. Dal Pozzolo et al. [36] noted that real-world deployment requires handling concept drift, where fraud patterns change over time as fraudsters adapt their strategies, and verification latency, where confirmed fraud labels arrive days or weeks after a transaction, leaving models to train on incomplete information [38].

Emerging Directions: Sequence Transformers, Federated Learning, and Large Language Models

Sequence transformers, federated learning, and large language models are recent additions to fraud detection research. Most are still evaluated on the same credit card datasets as established methods, and only federated learning is designed for the distributed, institution-held data that authorized transfer detection requires. Transformer architectures, which replaced recurrence with self-attention in sequence modeling [39], have begun to be applied to transaction streams. A unified transformer architecture combining multi-head self-attention with knowledge distillation, evaluated across the IEEE-CIS, European cardholder, and PaySim datasets, reported gains in both detection and interpretability over classical and deep baselines while remaining efficient enough for real-time scoring [33]. This extends the sequence modeling line already established by long short-term memory approaches [11-13], but it inherits the same benchmark concentration, since the evaluation datasets are again the public credit card and mobile payment corpora rather than any authorized transfer record.

Federated learning is the direction most directly relevant to the authorized-transfer gap. By training a shared model across institutions without moving raw data [40], federated frameworks target precisely the obstacle that keeps authorized transfer data out of public benchmarks. A federated transfer-learning framework designed for heterogeneous data distributions across banks reported improved recall over centralized state-of-the-art methods while keeping each institution's sample data in place and was explicitly positioned as a compliant cross-institutional risk-collaboration

paradigm [35]. The implication for this review is concrete: if authorized transfer detection is bottlenecked by data that cannot lawfully or competitively leave the institution that holds it, federated training is the candidate mechanism for building a model over that data without requiring a shared public dataset that does not, and may never, exist.

Large language models have been applied to fraud either by serializing tabular transactions into natural language representations or by reasoning over textual disclosures [34,41]. Their early results, however, recapitulate the metric problem that is central to this review rather than resolving it. An instruction-tuned large language model evaluated on the IEEE-CIS data reached 97.51 percent accuracy yet only 35.53 percent recall, far below the recall of an XGBoost baseline on the same data [34], a vivid reminder that a model can post a strong headline figure while catching barely a third of the fraud it is meant to stop. None of these three directions has yet produced a peer-reviewed detection-performance evaluation of authorized push payment fraud. Among them, federated learning is distinguished by a premise, modeling distributed, privacy-constrained institutional data, that matches the shape of the unmet problem more closely than the transformer and language model lines, which remain anchored to the same public credit card benchmarks as the methods they aim to surpass.

Explainability, Evaluation, and Regulatory Accountability

Explainability in fraud detection research is reported far more often than it is evaluated, with few studies testing whether the explanations they present are faithful to the underlying model. Post hoc methods predominate, and two approaches account for most applications. Lundberg and Lee [18] introduced SHAP, which assigns each feature an importance value for a particular prediction, grounded in Shapley values from cooperative game theory. Ribeiro et al. [19] introduced LIME, which explains the predictions of any classifier by learning an interpretable model locally around a prediction, operating in a model-agnostic manner by perturbing the input and observing how predictions change. Zafar and Wu [42] found SHAP and LIME to be the predominant explainability approaches in fraud detection research. Barredo Arrieta et al. [43] organized explainability methods into two taxonomies: one based on model transparency, encompassing simulatability, decomposability, and algorithmic transparency; and another based on post hoc techniques, including visualizations, local explanations, and feature relevance.

Lundberg and Lee [18] observed that the highest accuracy on large modern datasets is often achieved by complex ensemble or deep learning models that even experts struggle to interpret. Rudin [44] argued that black-box models are nonetheless deployed for high-stakes decisions and that attempting to explain them, rather than building inherently interpretable models, risks perpetuating poor practice. Rudin [44] further argued that, in auditable, high-stakes contexts, inherently interpretable models are preferable to post hoc explanations of black-box models, whose explanations may be unfaithful to the underlying model.

Zafar and Wu [42] reviewed 49 peer-reviewed studies published between 2021 and 2025 and found that approximately 80 percent used predictive model performance as a proxy for explanation quality rather than evaluating the explanations directly. The choice of metric matters. Mienye and Sun [20] noted that accuracy is misleading on imbalanced data, where a fitness function that weights majority and minority samples equally favors the majority class and returns high accuracy values that obscure poor minority detection. Du et al. [28] reported that the Matthews correlation coefficient is considered superior for binary classification because it accounts for all four confusion matrix categories rather than rewarding majority-class accuracy.

How class imbalance is addressed carries its own risks. Hayat and Magnier [45] identified four recurring methodological problems in the credit card fraud detection literature: data leakage resulting from preprocessing applied before the train-test split, vague reporting of methodological details, the absence of temporal validation for transaction data, and metric manipulation through recall optimization at the expense of precision [45]. Applying an oversampling step to the full dataset before splitting allows synthetic minority examples to leak into the test set. The same authors showed that a minimal neural network trained under this flawed procedure achieved 99.9 percent recall despite no genuine modeling advance [45]. This finding reframes the near-perfect figures commonly reported in the literature, including the cross-validated 0.9986 accuracy and perfect recall reported by Alarfaj et al. [31]: whether such results reflect genuine detection capability or evaluation leakage cannot be determined from the reported metrics alone. The choice of resampling

strategy further compounds the difficulty, as Hajek et al. [29] found that random under-sampling deteriorated classifier performance rather than improving it, contradicting the common assumption that rebalancing uniformly improves performance.

Zafar and Wu [42] found that common data resampling techniques used to manage class imbalance can distort background distributions and local neighborhoods, compromising the faithfulness and fidelity of post hoc explanations. They argued that intrinsically interpretable models remain preferable in auditable and high-risk contexts [42]. Iqbal et al. [46] combined random forest and XGBoost in an interpretable ensemble with SHAP and LIME explanations.

United States banking supervisors require financial institutions to manage model risk through validation, ongoing monitoring, and governance, including for machine learning models used in fraud detection [47]. The European Strong Customer Authentication requirement applies to unauthorized payment fraud [26]. The mandatory reimbursement requirement for authorized push payment fraud transfers the financial cost of undetected authorized fraud to the payment service provider [25].

Discussion

The clearest metric-bearing evidence in Table *1* remains concentrated in credit card benchmarks [7-8,10,28]. Across the reviewed literature, including recent systematic reviews spanning credit card, insurance, and financial statement fraud, the retained peer-reviewed sources contain no detection-performance evaluation of authorized push payment (APP) fraud, despite authorized push payment losses reaching 450.7 million pounds in the United Kingdom during 2024 [2] and bank transfers accounting for the largest share of United States fraud losses by payment method [1]. A single quantified APP figure surfaced, drawn from an industry pilot outside the peer-reviewed record, leaving the academic literature without a benchmark for this category. A structural reason is visible in the evidence landscape. An authorized payment initiated by the victim will ordinarily be completed because the instruction carries valid authorization [4], meaning that the anomaly detection and authentication mechanisms underlying the credit card methods have no irregular signal to detect. A model trained to distinguish fraudulent transactions from legitimate ones by their statistical fingerprint cannot flag a transaction that the legitimate account holder has authorized. This is not a tuning problem that a larger training set or a deeper network would resolve. Whatever signal the classifier learns to detect is absent by construction because the fraudulent intent resides with a third party whom the transaction record never captures. The evidence base and the loss data point in different directions: research concentrates where labeled transaction data are abundant and the fraud signal is detectable, while losses concentrate where the victim supplies the authorization. This concentration is self-reinforcing, as public credit card benchmarks invite subsequent studies to extend the same methods on the same data, while the authorized transfer problem, lacking any comparable public dataset, attracts none of that effort.

Performance figures that appear strong in isolation often weaken under deployment conditions. Mienye and Sun [20] reported a sensitivity of 0.997 on the European cardholders dataset but only 0.904 on the German dataset using the identical method, a 9-point difference attributable to dataset composition rather than method design. The European benchmark contains 284,807 transactions with a fraud rate of 0.172 percent [20,28], a distribution specific to one card network during a single month in 2013. Dal Pozzolo et al. [36] documented that real-world deployment introduces concept drift, whereby fraud patterns shift as fraudsters adapt, and verification latency, whereby confirmed labels arrive days or weeks after the transaction [38]. A model validated with an area under the curve of 0.990 on a static benchmark operates under conditions that the benchmark does not represent. Reported metrics therefore describe performance on a frozen snapshot rather than on the moving target that production systems face.

Dataset dependence is not the only reason the headline numbers warrant caution. Hayat and Magnier [45] argued that a substantial part of the credit card fraud literature is compromised by evaluation flaws rather than genuine modeling gains and demonstrated that a trivial neural network reached 99.9 percent recall purely because oversampling was applied before the train-test split, leaking synthetic fraud examples into the test data [45]. Read against that finding, the cross-validated 0.9986 accuracy and perfect recall reported by Alarfaj et al. [31] illustrate the interpretive problem rather than resolving it: a near-perfect score on this benchmark is as consistent with leakage as with detection skill, and the reported figures do not distinguish the two. The resampling step that produces these scores is itself contested because

Hajek et al. [29] found random under-sampling degraded rather than improved their mobile-payment classifiers [29], which means the imbalance correction so many studies treat as a routine preprocessing default can move performance in either direction depending on how and when it is applied. For a financial institution weighing which published method to adopt, this is the more consequential gap: not that methods are absent, but that the metrics distinguishing a strong method from a leaky evaluation are frequently absent.

Aggregate scores can also conceal the operational property that matters most to a financial institution. Du et al. [28] reported an autoencoder-XGBoost method with the highest Matthews correlation coefficient among four classifiers (0.8845), yet its true positive rate (0.7839) was the lowest of the four, trailing K-nearest neighbors at 0.8835 [28]. Here, the headline metric favored the weaker detector. A bank selecting a model on the basis of that metric would deploy the one that catches the fewest fraudulent transactions. The divergence between recall and precision further sharpens the point. Albalawi and Dardouri [7] reported that logistic regression achieved 90.32 percent recall but only 15.73 percent precision, meaning that roughly five out of every six flagged transactions were legitimate. In a deployment where each false positive blocks a genuine customer payment, such low precision renders the high recall operationally inadequate. The same divergence reappears in the newest family of methods: the instruction-tuned language model of Chen et al. [34] posted 97.51 percent accuracy but only 35.53 percent recall, showing that a state-of-the-art architecture and a textbook baseline can fail in the same direction, with a strong aggregate figure concealing weak fraud detection. Whichever metric a study chooses to foreground determines which method appears superior, and the metrics that flatter a model in a paper are not always those that govern its operational cost.

Read together, these findings trace a clear direction of travel and an equally clear stall. The trend across the strongest recent work is away from aggregate accuracy and toward evaluation that reflects deployment: cost-sensitive savings metrics that weight each transaction by its monetary value [9], interpretable ensembles that retain natural imbalance and report precision-recall rather than accuracy [32], and emerging transformer and federated designs that foreground interpretability and privacy [33,35]. Yet the literature also contains direct contradictions that any adopting institution must resolve for itself rather than inherit as settled. Resampling is treated as a routine default that improves minority detection [17], but at least one mobile-payment study found random under-sampling actively degraded performance [29], and resampling applied before splitting manufactures recall that survives no honest test [45]. Interpretability is presented as a property to be added post hoc through SHAP or LIME [18-19], yet the same resampling step that raises detection can distort the very explanations regulators would examine [42], and a competing position holds that auditable contexts call for intrinsically interpretable models in the first place [44]. The methods are not converging on a consensus; they are accumulating into a body of work whose internal disagreements are themselves the most useful finding for a practitioner, because they mark exactly where a published headline cannot be taken at face value.

Methods that improve detection performance can simultaneously undermine the explanations that regulation requires. Class imbalance necessitates resampling, and the Synthetic Minority Over-sampling Technique is the standard response [17]. Zafar and Wu [42] found that such resampling distorts the background distributions and local neighborhoods on which SHAP and LIME depend, compromising the fidelity of the resulting explanations. The same study also found that roughly 80 percent of explainable AI fraud studies never evaluate explanation quality directly, instead using model accuracy as a proxy [42]. This combination is consequential for regulated institutions. United States model risk management guidance requires the validation and ongoing monitoring of deployed models [47], yet the explanation tools on which most studies rely are both distorted by the imbalance correction those same studies apply and unvalidated against any direct measure of explanation quality. Rudin [44] argued that, in high-stakes contexts, inherently interpretable models are preferable to post hoc explanations of black-box models, a position that the evaluation gap in the fraud literature strengthens rather than resolves.

The United Kingdom's mandatory reimbursement requirement transfers the cost of undetected authorized push payment fraud to the payment service provider [25], making the detection of this specific fraud type a direct financial liability rather than merely a customer-service consideration. The peer-reviewed literature offers no validated method, no public benchmark, and no reported performance figure for addressing this liability. Strong Customer Authentication under the European framework reduces the unauthorized fraud that the methods in Table *1* already detect [26], leaving

the authorized category as the residual exposure. The institutions bearing the reimbursement cost hold the transaction data that public benchmarks lack, and federated learning offers a concrete approach to modeling these data across institutions without transferring them [35,40]. The most direct path toward closing this gap therefore lies in institution-held authorized transfer data, paired with explanation evaluation that measures fidelity rather than accuracy and is reported using metrics chosen for operational cost rather than headline performance.

This review has limitations. As a narrative rather than a fully systematic review, it did not register a protocol or perform dual independent screening, and its synthesis is qualitative rather than meta-analytic. Furthermore, the absence of any peer-reviewed study on authorized push payment fraud performance means that the central gap is established by the absence of evidence rather than by a contradictory body of results. The heavy reliance on credit card datasets reflects the availability of the underlying data rather than a selection preference, a constraint that the review treats as a finding in its own right.

Conclusions

Machine learning is mature for credit card fraud detection. It has barely reached authorized-transfer fraud, a category that mandatory reimbursement has turned into a direct cost for payment service providers. The strong results reported on credit card data do not transfer cleanly. They shift with the benchmark used, the metric a study chooses to report, and whether the evaluation guards against leakage, so that the same method can rank very differently across datasets and a high recall can sit alongside a precision too low to deploy. The explanation methods in common use are rarely validated, and the resampling that handles rare fraud can distort them while adding little to detection. Transformer and language-model methods have not escaped these problems, since they train on the same public data, and among the newer directions only federated learning suits data that must stay within each institution. The field does not need another benchmark result. It needs studies grounded in the authorized-transfer records institutions already hold, with evaluation that prevents leakage, weighs the cost of each decision, and checks that explanations hold up. The institutions now liable for these losses are also the ones holding the data, which puts them in a position to lead the work that the literature has so far left undone.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Kehinde Hamed, Moshood Sorinola

Acquisition, analysis, or interpretation of data: Kehinde Hamed, Moshood Sorinola

Drafting of the manuscript: Kehinde Hamed, Moshood Sorinola

Critical review of the manuscript for important intellectual content: Kehinde Hamed, Moshood Sorinola

Supervision: Moshood Sorinola

Disclosures

Conflicts of interest: In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

Data Availability Statements

Data sharing is not applicable. This work is theoretical; no datasets were generated or analyzed.

References

1. Federal Trade Commission: Consumer Sentinel Network Data Book 2024. (2025). Accessed: June 17, 2026: https://www.ftc.gov/reports/consumer-sentinel-network-data-book-2024.
2. UK Finance: Annual Fraud Report 2025. (2025). Accessed: June 17, 2026: https://www.ukfinance.org.uk/policy-and-guidance/reports-and-publications/annual-fraud-report-2025.
3. Nilson Report: Card Fraud Losses Worldwide — 2021. (2022). Accessed: June 17, 2026: https://nilsonreport.com/articles/card-fraud-losses-worldwide/.
4. Federal Reserve Bank of Kansas City: Payment Scams in Fast Payment Systems. (2024). Accessed: June 17, 2026: https://www.kansascityfed.org/research/payments-system-research-briefings/combating-authorized-push-payment-scams-in-....
5. Braithwaite J: 'Authorized push payment' bank fraud: what does an effective regulatory response look like?. Journal of Financial Regulation. 2024, 10:174-193. 10.1093/jfr/fjae006
6. West J, Bhattacharya M: Intelligent financial fraud detection: A comprehensive review. Computers & Security. 2016, 57:47-66. 10.1016/j.cose.2015.09.005
7. Albalawi T, Dardouri S: Enhancing credit card fraud detection using traditional and deep learning models with class imbalance mitigation. Frontiers in Artificial Intelligence. 2025, 8:1643292. 10.3389/frai.2025.1643292
8. Randhawa K, Loo CK, Seera M, Lim CP, Nandi AK: Credit card fraud detection using AdaBoost and majority voting. IEEE Access. 2018, 6:14277-14284. 10.1109/access.2018.2806420
9. Almhaithawi D, Jafar A, Aljnidi M: Example-dependent cost-sensitive credit cards fraud detection using SMOTE and Bayes minimum risk. SN Applied Sciences. 2020, 2:1574. 10.1007/s42452-020-03375-w
10. Mienye ID, Jere N: Deep learning for credit card fraud detection: a review of algorithms, challenges, and solutions. IEEE Access. 2024, 12:96893-96910. 10.1109/access.2024.3426955
11. Benchaji I, Douzi S, El Ouahidi B, Jaafari J: Enhanced credit card fraud detection based on attention mechanism and LSTM deep model. Journal of Big Data. 2021, 8:151. 10.1186/s40537-021-00541-8
12. Jurgovsky J, Granitzer M, Ziegler K, Calabretto S, Portier PE, He-Guelton L, Caelen O: Sequence classification for credit-card fraud detection. Expert Systems with Applications. 2018, 100:234-245. 10.1016/j.eswa.2018.01.037
13. Esenogho E, Mienye ID, Swart TG, Aruleba K, Obaido G: A neural network ensemble with feature engineering for improved credit card fraud detection. IEEE Access. 2022, 10:16400-16407. 10.1109/access.2022.3148298
14. Fiore U, De Santis A, Perla F, Zanetti P, Palmieri F: Using generative adversarial networks for improving classification effectiveness in credit card fraud detection. Information Sciences. 2019, 479:448-455. 10.1016/j.ins.2017.12.030
15. Liu FT, Ting KM, Zhou ZH: Isolation forest. 2008 Eighth IEEE International Conference on Data Mining, Pisa, Italy. 2008, 413-422. 10.1109/ICDM.2008.17
16. Motie S, Raahemi B: Financial fraud detection using graph neural networks: A systematic review. Expert Systems with Applications. 2024, 240:122156. 10.1016/j.eswa.2023.122156
17. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP: SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research. 2002, 16:321-357. 10.1613/jair.953
18. Lundberg SM, Lee S-I: A unified approach to interpreting model predictions. NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017, 4768-4777.
19. Ribeiro MT, Singh S, Guestrin C: "Why Should I Trust You?": Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016, 1135-1144. 10.1145/2939672.2939778
20. Mienye ID, Sun YA: A machine learning method with hybrid feature selection for improved credit card fraud detection. Applied Sciences. 2023, 13:7254. 10.3390/app13127254

How to cite this article:
Sorinola M, Hamed K (July 03, 2026) Machine Learning for Digital Payment Fraud Detection in Financial Institutions: A Review of Methods, Explainability, and Performance. Cureus J Comput Sci 3 : es44389-026-00149-0. DOI https://doi.org/10.7759/s44389-026-00149-0

21. Ali A, Abd Razak S, Othman SH, et al.: Financial fraud detection based on machine learning: A systematic literature review. Applied Sciences. 2022, 12:9637. 10.3390/app12199637

22. Hilal W, Gadsden SA, Yawney J: Financial fraud: A review of anomaly detection techniques and recent advances. Expert Systems with Applications. 2022, 193:116429. 10.1016/j.eswa.2021.116429

23. Cherif A, Badhib A, Ammar H, Alshehri S, Kalkatawi M, Imine A: Credit card fraud detection in the era of disruptive technologies: A systematic review. Journal of King Saud University - Computer and Information Sciences. 2023, 35:145-174. 10.1016/j.jksuci.2022.11.008

24. Compagnino AA, Maruccia Y, Cavuoti S, Riccio G, Tutone A, Crupi R, Pagliaro A: An Introduction to Machine Learning Methods for Fraud Detection. Applied Sciences. 2025, 15:11787. 10.3390/app152111787

25. Payment Systems Regulator: PS23/3: Fighting authorised push payment fraud: a new reimbursement requirement. (2023). Accessed: June 17, 2026: https://www.psr.org.uk/publications/policy-statements/ps233-fighting-authorised-push-payment-fraud-a-new-reimbursemen....

26. European Parliament and Council of the European Union: Directive (EU) 2015/2366 on payment services in the internal market (PSD2). (2015). Accessed: June 17, 2026: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32015L2366.

27. Tayebi M, El Kafhali S: A novel approach based on XGBoost classifier and Bayesian optimization for credit card fraud detection. Cyber Security and Applications. 2025, 3:100093. 10.1016/j.csa.2025.100093

28. Du H, Lv L, Wang H, Guo A: A novel method for detecting credit card fraud problems. PLoS ONE. 2024, 19:e0294537. 10.1371/journal.pone.0294537

29. Hajek P, Abedin MZ, Sivarajah U: Fraud detection in mobile payment systems using an XGBoost-based framework. Information Systems Frontiers. 2023, 25:1985-2003. 10.1007/s10796-022-10346-6

30. Almazroi AA, Ayub N: Online payment fraud detection model using machine learning techniques. IEEE Access. 2023, 11:137188-137203. 10.1109/access.2023.3339226

31. Alarfaj FK, Malik I, Khan HU, Almusallam N, Ramzan M, Ahmed M: Credit card fraud detection using state-of-the-art machine learning and deep learning algorithms. IEEE Access. 2022, 10:39700-39715. 10.1109/access.2022.3166891

32. Baisholan N, Dietz JE, Gnatyuk S, Turdalyuly M, Matson ET, Baisholanova K: FraudX AI: An interpretable machine learning framework for credit card fraud detection on imbalanced datasets. Computers. 2025, 14:120. 10.3390/computers14040120

33. Dubey P, Dubey P, Bokoro PN: A unified transformer-BDI architecture for financial fraud detection: Distributed knowledge transfer across diverse datasets. Forecasting. 2025, 7:31. 10.3390/forecast7020031

34. Chen J, Quintano J, Qiang Y: FinFraud-LLM: Exploring large language models for financial fraud detection. 2025 IEEE International Conference on Data Mining Workshops (ICDMW), Washington, DC, USA. 2025, 3087-3091. 10.1109/ICDMW69685.2025.00407

35. Chen Y, Zhang K, Zhu H, Qiu Z: A novel federated transfer learning framework for credit card fraud detection under heterogeneous data conditions. Risks. 2025, 13:208. 10.3390/risks13110208

36. Dal Pozzolo A, Caelen O, Le Borgne YA, Waterschoot S, Bontempi G: Learned lessons in credit card fraud detection from a practitioner perspective. Expert Systems with Applications. 2014, 41:4915-4928. 10.1016/j.eswa.2014.02.026

37. Bahnsen AC, Aouada D, Stojanovic A, Ottersten B: Feature engineering strategies for credit card fraud detection. Expert Systems with Applications. 2016, 51:134-142. 10.1016/j.eswa.2015.12.030

38. Dal Pozzolo A, Boracchi G, Caelen O, Alippi C, Bontempi G: Credit card fraud detection: a realistic modeling and a novel learning strategy. IEEE Transactions on Neural Networks and Learning Systems. 2018, 29:3784-3797. 10.1109/tnnls.2017.2736643

39. Vaswani A, Shazeer N, Parmar N, et al.: Attention is all you need. NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017, 30:6000-6010.

40. McMahan B, Moore E, Ramage D, Hampson S, Agüera y Arcas B: Communication-efficient learning of deep networks from decentralized data. Proceedings of the 20th International Conference on Artificial Intelligence and Statistics. 2017, 54:1273-1282.

41. Dai Y, Qian Y, Wang X, Liu Y, Wang C: FSFDLLM: Financial statement fraud detection aided by large language models. Intelligent Computing. 2026, 5:0274. 10.34133/icomputing.0274

42. Zafar U, Wu F: Methodological challenges in explainable AI for fraud detection: a systematic literature review. Artificial Intelligence Review. 2026, 59:115. 10.1007/s10462-026-11516-7

How to cite this article:
Sorinola M, Hamed K (July 03, 2026) Machine Learning for Digital Payment Fraud Detection in Financial Institutions: A Review of Methods, Explainability, and Performance. Cureus J Comput Sci 3 : es44389-026-00149-0. DOI https://doi.org/10.7759/s44389-026-00149-0

43. Barredo Arrieta A, Diaz-Rodriguez N, Del Ser J, et al.: Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion. 2020, 58:82-115. 10.1016/j.inffus.2019.12.012

44. Rudin C: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence. 2019, 1:206-215. 10.1038/s42256-019-0048-x

45. Hayat K, Magnier B: Data leakage and deceptive performance: A critical examination of credit card fraud detection methodologies. Mathematics. 2025, 13:2563. 10.3390/math13162563

46. Iqbal S, Awan KM, Kamal S, Rehman ZU: Interpretable ensemble learning models for credit card fraud detection. Applied Sciences. 2025, 15:12073. 10.3390/app152212073

47. Federal Deposit Insurance Corporation, Office of the Comptroller of the Currency: Revised Guidance on Model Risk Management (SR 26-2). (2026). Accessed: June 17, 2026: https://www.federalreserve.gov/supervisionreg/srletters/SR2602.pdf.

How to cite this article:
Sorinola M, Hamed K (July 03, 2026) Machine Learning for Digital Payment Fraud Detection in Financial Institutions: A Review of Methods, Explainability, and Performance. Cureus J Comput Sci 3 : es44389-026-00149-0. DOI https://doi.org/10.7759/s44389-026-00149-0