Vijay Saini, 2026, 14:3
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

Explainable Credit Card Fraud Detection Using LightGBM and SHAP-Guided Feature Selection: A Review

Vijay Saini, Jitender Kumar
Department of Computer Science and Engineering, Ganga Institute of Technology and Management, Kablana, Jhajjar, Haryana, India,

Abstract- Digital payment ecosystems have expanded both the volume and complexity of transaction data, widening the opportunity for fraud while raising customer, operational, and regulatory expectations that automated decisions remain explainable. This review consolidates research on credit card fraud detection spanning rule-based systems, classical statistical learning, ensemble and gradient-boosted methods, deep learning architectures, and explainable artificial intelligence (XAI), and synthesises these strands into FraudDetectNet, a layered pipeline in which data preprocessing, feature reduction, classification, explanation generation, and monitoring are treated as interdependent rather than separable functions. Particular attention is given to Shapley Additive Explanations (SHAP), used here not only as a post-hoc diagnostic but as an upstream feature-selection mechanism that compresses high-dimensional transaction data while preserving discriminative power. The review also examines evaluation practice for severely imbalanced, temporally ordered fraud data, arguing for precision-recall and cost-sensitive measures over plain accuracy. Outcomes reported in the primary reference study underlying this review — feature-space reduction from 380 to 120 variables, 97.6% accuracy, 99.0% fraud-class recall, and training-time reduction from 185.3 to 54.1 seconds — are presented as evidence of feasibility rather than generalised guarantees, and extensions toward graph-based, federated, and continual learning are outlined.

Keywords: credit card fraud detection; explainable artificial intelligence; SHAP; LightGBM; class imbalance; concept drift; gradient boosting.

I. INTRODUCTION

The growth of digital commerce has increased both the volume and heterogeneity of transaction data, creating economic opportunity alongside a widening attack surface for fraud. A fraud score that cannot be justified, monitored, and reviewed carries institutional risk comparable to a weak classifier: false negatives translate into direct monetary loss, while false positives generate customer friction, lost authorisations, and analyst workload. Traditional rule-based systems remain transparent but are brittle against evolving fraud behaviour, whereas many high-performing deep learning models remain difficult to justify in regulated financial environments. Consequently, fraud detection research has shifted toward workflows that combine robust preprocessing, gradient-boosted ensembles, principled handling of skewed classes, and explanation-aware governance, rather than optimising predictive accuracy in isolation.

This review is organised around a primary reference study that applies SHAP-guided feature pruning to a LightGBM-based fraud classifier trained on the IEEE-CIS transaction dataset, using explanation signals to simplify the model before deployment rather than only auditing it afterward. Building on this foundation, the review (i) surveys the methodological evolution of fraud detection from rule-based systems to explainable AI, (ii) formalises the operational and statistical constraints that a production-grade fraud classifier must satisfy, (iii) describes FraudDetectNet, an explainable architecture that generalises the reference study's approach into a modular pipeline, and (iv) proposes an evaluation strategy suited to imbalanced, temporally ordered, and explanation-sensitive fraud data.

The remainder of the paper is organised as follows. Section II reviews the relevant literature across five methodological families. Section III formalises the fraud detection problem and the constraints that

motivate the proposed architecture. Section IV describes the FraudDetectNet pipeline and its SHAP-guided feature selection engine. Section V outlines the evaluation strategy and discusses the primary study's reported outcomes alongside their limitations. Section VI concludes and identifies directions for future work.

II. LITERATURE REVIEW

A. Evolution of Fraud Detection Methods

Fraud screening has progressed through five broad phases: manual rules and expert heuristics, statistical and classical machine learning classifiers, ensemble and gradient-boosted tree methods, deep learning architectures, and, more recently, explainable AI approaches that attempt to recover the interpretability lost in the transition to ensemble and deep models. Each phase has generally traded some transparency for predictive strength: hand-crafted rules are easy to audit but brittle against adversarial adaptation, while deep architectures capture complex non-linear patterns at the cost of opacity. This trajectory motivates treating explainability not as a final audit step but as a design constraint integrated from the outset — the premise examined throughout this review.

B. Rule-Based, Statistical, and Classical Machine Learning Approaches

Rule-based systems remain valuable for their directness: they encode policy explicitly, support immediate intervention, and are straightforward to audit, but they struggle to keep pace with evolving fraud tactics and require continuous manual maintenance. Logistic regression, decision trees, and support vector machines extended this transparency into a statistical setting, offering interpretable coefficients and a useful baseline, but are limited in capturing the non-linear feature interactions that characterise sophisticated fraud. Bagging ensembles such as random forests improved robustness to noisy, mixed-type features but introduced additional latency and only partial interpretability. These approaches remain useful as fallback controls and baseline comparators rather than as primary detection engines for large-scale transaction streams, and they inform the complementary rule layer and baseline comparators used in FraudDetectNet rather than its core learner.

C. Ensemble Learning and Gradient Boosting

Gradient boosting has become a dominant paradigm for large-scale tabular fraud detection because it captures non-linear interactions while remaining computationally more tractable than many deep architectures. LightGBM, in particular, uses leaf-wise tree growth together with Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) to achieve strong predictive performance with efficient training, making it well suited to high-dimensional, high-volume transaction data. XGBoost offers comparable strengths through a different regularisation and split-finding strategy. Both frameworks, however, remain opaque without an explicit explanation layer, since their ensemble structure obscures the basis for individual predictions. This tension between strong tabular performance and residual opacity is what the primary reference study addresses by integrating SHAP attribution directly into the LightGBM training pipeline, motivating LightGBM's selection as FraudDetectNet's core learner. Table I summarises the comparative strengths and limitations of the modelling families considered in this review.

TABLE I. COMPARATIVE OVERVIEW OF MAJOR FRAUD DETECTION MODELLING FAMILIES

Approach	Strength	Limitation	Relevance to FraudDetectNet
Rule-based systems	Transparent, policy-aligned	Brittle against evolving fraud	Fallback rules / business constraints
Statistical models (LR, SVM)	Simple, auditable	Weak on non-linear interactions	Baseline comparator
Bagging ensembles (RF)	Stable on mixed features	Added latency, partial interpretability	Comparative benchmark

LightGBM / XGBoost	Strong tabular performance	Opaque without explanation layer	Core modelling engine
Deep learning (ANN/AE/RNN/GNN)	Learns complex / relational patterns	High compute cost, low interpretability	Comparator and future extension
Explainable AI (SHAP, LIME)	Improves trust and auditability	Can remain post-hoc only	Integrated upstream in feature selection

D. Deep Learning Architectures for Fraud Detection

Deep neural networks, autoencoders, recurrent networks, and graph neural networks offer expressive modelling capacity beyond gradient-boosted trees, particularly for sequential or relational data. Artificial neural networks learn complex non-linear relationships but require careful tuning and an additional explanation mechanism to be usable in regulated settings. Autoencoders are attractive for unsupervised or semi-supervised detection of rare and previously unseen fraud patterns, though their reconstruction-error thresholds can be unstable. Recurrent architectures model the sequential structure of a customer's transaction history but incur higher computational cost than tree-based methods. Graph neural networks capture relationships among cards, merchants, devices, and IP addresses, making them well suited to detecting coordinated fraud rings, but require explicit graph construction and are correspondingly heavier to deploy within authorisation-time latency budgets. These trade-offs motivate using deep learning primarily as a comparator and future extension within FraudDetectNet rather than as its production core.

E. Explainable AI in Financial Decision Systems

Explainable AI (XAI) research addresses how institutions justify automated decisions to analysts, auditors, customers, and regulators. SHAP provides both global feature-contribution rankings and local, transaction-level explanations grounded in cooperative game theory, and is widely regarded as a consistent model-agnostic attribution method for tabular data. LIME offers local surrogate explanations that are computationally lighter but less stable across similar transactions. Global feature-importance rankings support model governance and periodic feature review, while counterfactual explanations support contestability for declined or flagged transactions, and reason codes translate attributions into human-readable statements for case management and audit trails. The distinguishing contribution of the primary reference study, carried into FraudDetectNet, is using SHAP attribution upstream — as a feature-selection criterion before final model training — rather than purely as a post-hoc diagnostic applied after a model is fixed. This reframes explainability as a mechanism for model compression and clarity, not only retrospective justification.

F. Class Imbalance, Cost Asymmetry, and Temporal Drift

Because confirmed fraud cases typically represent a small fraction of all transactions, naive optimisation toward accuracy rewards majority-class prediction while missing the minority-class events that matter operationally. The literature addresses this through resampling techniques such as SMOTE, cost-sensitive loss functions, threshold tuning, and minority-focused metrics including the precision-recall area under the curve (PR-AUC) and the Matthews correlation coefficient (MCC), which are more informative than accuracy under severe skew. A related challenge is non-stationarity: fraud tactics, spending behaviour, and merchant risk profiles evolve over time, so studies relying on random data splits risk overestimating real-world performance. Chronological train-validation-test partitioning, ongoing drift monitoring, and scheduled or triggered retraining are therefore treated as evaluation requirements rather than optional refinements; adaptive window-based drift detection has also been examined as a mechanism for identifying when retraining is warranted. Imbalance handling and temporal validation together define the evaluation discipline adopted in Section V.

G. Literature Gaps and Synthesis

Across these strands, predictive modelling has advanced substantially, but a persistent gap remains between high accuracy and institutionally credible explanation: many models remain effectively black-box in production, explainability is frequently treated as a post-hoc add-on rather than a design objective, and few studies integrate feature pruning, explanation generation, and governance monitoring into a single, coherent pipeline. This gap motivates the FraudDetectNet architecture described in Section IV, which positions SHAP-guided feature selection, model training, explanation delivery, and drift monitoring as co-dependent components of one explainable analytics pipeline rather than as separable research threads.

III. PROBLEM FORMULATION

At its core, credit card fraud detection assigns an incoming transaction, represented by a feature vector x, a probability $p(y=1|x)$ that it is fraudulent, where y = 1 denotes fraud and y = 0 denotes a genuine transaction. A decision threshold T converts this probability into an action, and the resulting confusion matrix — true positives, false positives, false negatives, and true negatives — carries asymmetric business consequences: a false negative may permit direct monetary loss and reputational damage, whereas a false positive inconveniences a genuine customer, reduces authorisation rates, and adds to manual-review workload. A rigorous formulation must therefore treat the problem as constrained optimisation rather than single-metric classification.

Table II summarises the principal constraints that distinguish production fraud detection from a conventional supervised-learning benchmark, together with the design response adopted in FraudDetectNet. Beyond accuracy, a deployable system must (i) maximise minority-class detection without inflating the false-alarm burden carried by human reviewers, (ii) operate within authorisation-time latency budgets, typically on the order of milliseconds, and (iii) produce explanations and audit logs usable by analysts and regulators. Framed formally, the objective is to learn a classifier that maximises a cost-sensitive utility function subject to a latency bound and an explanation-fidelity requirement, rather than maximising raw discriminative accuracy alone.

TABLE II. KEY CONSTRAINTS IN PRODUCTION FRAUD DETECTION AND THE FRAUDDETECTNET DESIGN RESPONSE

Constraint	Formal Meaning	Design Response
Class imbalance	Fraud cases are rare relative to genuine transactions	PR-AUC, recall, MCC, threshold tuning
Cost asymmetry	False negatives and false positives carry different costs	Cost-sensitive loss and calibrated thresholds
Temporal drift	Fraud patterns change over time	Chronological validation and monitoring
Explainability	Decisions require justification	SHAP, reason codes, audit logs
Latency	Authorisation decisions must be fast	Feature pruning and scoring-time monitoring

This formulation yields four propositions examined throughout the review: (P1) explainability-guided feature selection can reduce feature-space redundancy while preserving minority-class discrimination; (P2) temporal validation provides a more realistic estimate of deployed performance than random splitting; (P3) a pipeline that logs local explanations improves analyst trust and audit readiness relative to an equally accurate opaque model; and (P4) a modular architecture combining feature engineering, SHAP-guided pruning, and LightGBM can better balance latency and transparency than heavier deep learning alternatives.

IV. PROPOSED FRAUDDETECTNET FRAMEWORK

Building on the constraints identified in Section III, FraudDetectNet is conceived as a layered explainable analytics pipeline in which preprocessing, feature reduction, classification, and explanation are treated as co-dependent rather than sequential, disconnected stages. The pipeline

comprises seven functional layers. The data ingestion layer consolidates transaction, identity, device, and merchant signals into a unified event schema with consistent timestamps and entity identifiers, supporting both batch and incremental processing. The preprocessing layer performs missing-value handling, categorical encoding — favouring target encoding for high-cardinality fields such as merchant or device identifiers and one-hot encoding for low-cardinality categories — and chronological train-validation-test partitioning to avoid look-ahead bias. The feature engineering layer constructs behavioural aggregates such as transaction velocity, amount deviation, and device-merchant interaction signals, which the primary reference study identifies as especially salient predictors under SHAP analysis.

The SHAP-guided feature selection engine forms the central novelty of the architecture. A baseline model is trained on the retained feature pool; mean absolute SHAP values are computed and used to rank global feature importance; low-signal features are eliminated against a chosen threshold; and the model is retrained on the reduced set, with the loop repeated until performance stabilises. This compresses the candidate feature space considerably before the LightGBM training core — configured with leaf-wise growth, GOSS/EFB acceleration, class weighting via scale_pos_weight, and hyperparameter search — produces a calibrated fraud probability for each transaction.

The local explainability layer then generates SHAP-based reason codes for every scored transaction, identifying which features pushed the decision toward or away from the fraud label, and packages these as audit-ready case attachments. The inference layer maps the resulting probability into a risk-tiered action: low-probability transactions are approved without added friction, medium-probability transactions trigger step-up authentication, and high-probability transactions are routed to manual review or declined, with explanation logging at every tier. Finally, the monitoring and governance layer tracks feature drift, segment-level fairness, p95 latency, PR-AUC and recall against accepted thresholds, and explanation stability, triggering scheduled or event-based retraining when degradation is detected. Security and access controls — role-based access to explanations, pseudonymisation, and controlled audit retrieval — are treated as cross-cutting requirements across all seven layers, since explainability must not become a vector for overexposing sensitive customer data.

V. EVALUATION STRATEGY AND REPORTED OUTCOMES

Evaluating a fraud detection system under severe class imbalance requires moving beyond accuracy toward metrics that reflect minority-class performance and operational cost. Table III summarises the metrics adopted in this review: precision and recall characterise alert purity and missed-fraud risk respectively; the F1 score balances the two; PR-AUC is preferred over AUC-ROC under extreme skew because it is more sensitive to minority-class performance; MCC offers a balanced correlation measure robust to imbalance; and latency together with cost-sensitive loss connects statistical performance to deployability and business impact.

TABLE III. EVALUATION METRICS FOR IMBALANCED FRAUD DETECTION

Metric	Why It Matters	Interpretation
Precision	Alert purity	Reduces unnecessary friction and analyst overload
Recall	Missed-fraud detection	Critical, since false negatives cause direct loss
F1 Score	Precision-recall balance	Useful when capture and alert efficiency both matter
PR-AUC	Minority-class ranking quality	More informative than accuracy / AUC-ROC under skew
MCC	Balanced correlation	Robust under severe class imbalance
Latency & cost-sensitive loss	Deployability and business impact	Supports authorisation-time decisions and ROI assessment

A credible evaluation protocol benchmarks the proposed SHAP-pruned LightGBM model against simpler and competing approaches — logistic regression, random forest, unpruned LightGBM/XGBoost, and a deep learning comparator such as an ANN — so that observed gains can be attributed to the SHAP-guided design rather than dataset idiosyncrasies. Ablation studies comparing full-feature against reduced-feature variants, and varying the SHAP elimination threshold, are necessary to isolate the contribution of feature pruning itself. Explainability is evaluated along complementary criteria: faithfulness (whether explanations reflect the model's actual decision logic, verified through consistency and perturbation checks), stability (whether similar transactions receive similar explanations), usefulness (whether analysts can act on generated reason codes), and auditability (whether explanations can be stored and retrieved for later review).

The primary reference study, conducted on the IEEE-CIS Fraud Detection dataset using a chronological train-validation-test split, reports that SHAP-guided pruning reduced the candidate feature set from 380 to 120 variables while preserving strong discrimination: 97.6% accuracy, 99.0% recall on the fraud class, and an AUC-ROC of 0.982. Training latency fell from 185.3 to 54.1 seconds, and inference latency remained under two milliseconds per transaction — a result directly relevant to authorisation-time decision support. These figures are treated in this review as benchmark evidence of feasibility on one dataset rather than as a generalised guarantee, since dataset age, possible label noise, and differences between laboratory and live payment traffic constitute genuine threats to external validity. A responsible deployment programme would still require institution-specific recalibration, threshold tuning, fairness auditing, and continuous post-deployment monitoring before the reported gains could be assumed to transfer.

VI. CONCLUSION AND FUTURE SCOPE

This review has argued that predictive performance alone is insufficient for high-stakes financial fraud detection: a deployable system must also be interpretable, auditable, efficient, and robust to changing transaction behaviour. Synthesising the literature on rule-based systems, statistical and ensemble learning, deep learning, and explainable AI, the review consolidates these requirements into FraudDetectNet, an architecture in which SHAP-guided feature selection, LightGBM-based classification, local and global explanation, and continuous monitoring are designed together rather than bolted on sequentially. The evidence reviewed from the primary reference study supports the proposition that explainability-guided feature reduction can simultaneously improve training efficiency and preserve — rather than sacrifice — predictive performance, reframing SHAP as a design-time compression tool and not merely a retrospective audit artefact.

Several directions merit further research. Graph neural networks could extend the architecture to model relationships among cards, merchants, devices, and IP addresses, improving detection of coordinated fraud rings. Federated learning would allow cross-institution model training without sharing raw customer data, addressing a key privacy constraint in financial collaboration. Continual learning under rolling time windows could address concept drift more directly than periodic retraining alone, while human-in-the-loop case review systems could use analyst feedback to refine both model behaviour and explanation quality. Finally, embedding model cards, fairness checks, and audit trails as first-class architectural components — rather than external compliance documentation — would strengthen the readiness of explainable fraud detection systems for regulatory scrutiny.

REFERENCES

1. V. Saini and J. Kumar, "Explainable Fraud Detection with LightGBM on Large-Scale Transaction Data via SHAP-Guided Feature Selection," primary reference manuscript.
2. N. Bussmann, P. Giudici, D. Marinelli, and J. Papenbrock, "Explainable machine learning in credit risk management," Computational Economics, vol. 57, pp. 203–216, 2021.

3. M. Bücker, G. Szepannek, A. Gosiewska, and P. Biecek, "Transparency, auditability, and explainability of machine learning models in credit scoring," Journal of the Operational Research Society, vol. 73, no. 1, pp. 70–90, 2022.
4. I. Almubark, "Advanced Credit Card Fraud Detection: An Ensemble Learning Using Random Under Sampling and Two-Stage Thresholding," IEEE Access, vol. 12, pp. 34120–34135, 2024.
5. E. Esenogho, I. D. Mienye, T. G. Swart, K. Aruleba, and G. Obaido, "A neural network ensemble with feature engineering for improved credit card fraud detection," IEEE Access, vol. 10, pp. 16400–16407, 2022.
6. P. Thanathamathee, S. Sawangarreerak, S. Chantamunee, and D. N. Mohd Nizam, "Robust and explainable financial fraud detection utilizing gradient boosting frameworks," Emerging Science Journal, vol. 8, no. 6, pp. 1120–1135, 2024.
7. S. Yusuf and D. Ekrem, "Detecting Credit Card Fraud by ANN and Logistic Regression," in Proc. Int. Symp. Innovations in Intelligent Systems and Applications, 2011.
8. K. S. Lim, L. H. Lee, and Y. W. Sim, "A Review of Machine Learning Algorithms for Fraud Detection in Credit Card Transaction," International Journal of Computer Applications, vol. 182, no. 42, pp. 12–21, 2019.
9. F. D. L. Bourdonnaye and F. Daniel, "Evaluating categorical encoding methods on a real credit card fraud detection database," arXiv:2112.12024, 2021.
10. J. Chen, X. Wang, and Y. Zhang, "Deep Learning in Financial Fraud Detection: Innovations, Challenges, and Applications," Data Science and Management, vol. 8, no. 3, pp. 210–225, 2025.
11. A. V. Ponce-Bobadilla, V. Schmitt, C. S. Maier, S. Mensing, and S. Stodtmann, "Practical guide to SHAP analysis: explaining supervised machine learning model predictions in drug development," Clinical and Translational Science, vol. 17, no. 4, e70056, 2024.
12. T. Altyeb, M. Ahmed, and S. Othman, "Bayesian hyperparameter optimization algorithm for tuning LightGBM parameters in banking pipelines," CEUR Workshop Proceedings, vol. 3478, pp. 62–74, 2023.
13. M. P. Da Silva, "Explainable AI and feature selection: Performance gains using Shapley additive attributions," Bachelor's thesis, University of Brasília, 2021.
14. Y. Ding and L. Byrne, "Risk-Informed Machine Learning Models for Renewal Classification in Motor Insurance using SHAP Frameworks," Journal of Risk and Financial Management, vol. 18, no. 1, pp. 45–58, 2025.
15. X. Cai and Z. Xie, "Financial Statement Fraud Detection Through an Integrated Machine Learning and Explainable AI Framework," IEEE Transactions on Information Forensics and Security, vol. 19, pp. 1204–1218, 2024.
16. L. M. Valle-Cruz and J. R. Gil-Garcia, "An Explainable AI Framework for Fraud Detection in Financial Ecosystems: Balancing Accuracy, Interpretability, and Compliance," Government Information Quarterly, vol. 42, no. 1, 101920, 2025.
17. S. Caxton Emerald and T. Ananthakrishnan, "Concept Drift Handling in Continuous Fraud Detection Systems via ADWIN Window Aggregations," IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 54, no. 3, pp. 1720–1733, 2024.
18. Y. Yang, L. M. Zhang, and Q. T. Zhou, "Federated learning for credit card fraud detection with data balancing techniques," Journal of Network and Computer Applications, vol. 221, 103729, 2024.
19. W. Walauskis, J. Rosen, J. Russell, and P. Kartik, "SHAP-based Feature Selection for Enhanced Unsupervised Labeling on Highly Imbalanced Data," Informatics in Medicine Unlocked, vol. 30, 100924, 2022.
20. G. Sun, M. Mirmehdi, Z. Abdallah, R. Santos-Rodriguez, I. Craddock, and T. M. S. Filho, "TACTFL: Temporal Contrastive Training for Multi-modal Federated Learning with Similarity-guided Model Aggregation," in Proc. 36th British Machine Vision Conference, 2025.