MSRDG International Journal of Computer Scientific Technology & Electronics Engineering ISSN: 3107 - 426X
Volume 2 Issue 1, 27-32, February 2026

Machine Learning Strategies for Securing Financial Transactions against Risks

Pusa Prashanth[1], Sathwika Gade[2]

[1]*Department of Data Science and Artificial Intelligence, Sheffield Hallam University, Sheffield, London, UK.*
[2]*Department of Data Science and Computational Intelligence, Coventry University, Coventry, West Midlands, United Kingdom.*
[1]*c5055216@hallam.shu.ac.uk*

Received: 05.01.2026 Revised:06.02.20256 Accepted: 14.02.2026 Published:28.02.2026

Abstract - *Financial institutions worldwide face escalating threats from sophisticated fraudsters who continuously adapt their methods to circumvent conventional rule-based detection systems. This paper introduces a multi-layer machine learning framework designed to identify and neutralise fraudulent financial transactions with high precision and minimal false-alarm rates. Our approach integrates an ensemble of Random Forest, Extreme Gradient Boosting (XGBoost), Long Short-Term Memory (LSTM) networks, and Isolation Forest algorithms, unified through a soft-voting fusion mechanism. We address class imbalance — a pervasive challenge in fraud datasets — using a hybrid Synthetic Minority Over-sampling Technique (SMOTE) combined with edited nearest-neighbour undersampling. Feature engineering incorporates temporal transaction patterns, geospatial anomaly indicators, device-fingerprinting scores, and merchant-level risk ratings derived from historical behavioural analytics. Extensive experiments conducted on two publicly available datasets (IEEE-CIS Fraud Detection and PaySim) and one proprietary bank dataset (n = 1.2 million records) demonstrate that the proposed ensemble achieves an accuracy of 98.7%, an AUC-ROC of 0.994, a precision of 98.1%, recall of 97.9%, and an F1-score of 98.0%, outperforming six baseline models. Ablation studies confirm the complementary contribution of each sub-model. An explainability layer using SHAP (SHapley Additive exPlanations) values renders the model decisions interpretable for compliance officers. The framework is deployable in real-time streaming environments with a mean inference latency of 12.4 ms per transaction, satisfying the latency constraints of production payment systems.*

Keywords - *Fraud detection · Machine learning · XGBoost · LSTM · Ensemble methods · Class imbalance · SMOTE · Financial risk management · SHAP interpretability · Real-time inference*

1. Introduction

The global digitalisation of financial services has generated unprecedented convenience for consumers but has simultaneously created fertile ground for novel categories of financial crime. According to the Association of Certified Fraud Examiners, organisations worldwide lose approximately 5% of their annual revenues to fraud, with financial services remaining among the most targeted sectors [1]. The proliferation of contactless payments, mobile banking applications, cross-border cryptocurrency transfers, and buy-now-pay-later schemes has dramatically expanded the attack surface available to malicious actors. Traditional rule-based fraud detection systems, which rely on expert-curated threshold conditions, are increasingly unable to capture the nuanced, adaptive behaviours of modern fraudsters who deliberately design transaction sequences to remain below detection thresholds.

Machine learning (ML) offers a paradigm shift: rather than encoding brittle human-defined rules, ML models infer complex, high-dimensional decision boundaries directly from historical labelled transaction data. Early applications of ML to fraud detection employed logistic regression and decision trees [2, 3]. Subsequent work demonstrated the superiority of ensemble methods — particularly Random Forests and gradient boosting variants — in capturing non-linear feature interactions [4, 5]. The advent of deep learning introduced the possibility of exploiting temporal sequences of transactions through recurrent architectures such as LSTM networks [6, 7]. More recently, “graph neural networks” (GNNs) have been applied to model transactional relationship graphs [8], and transformer-based architectures have been adapted for tabular financial data [9].

Despite this rich literature, several challenges remain incompletely addressed. First, extreme class imbalance — legitimate transactions typically outnumber fraudulent ones by ratios exceeding 500:1 — causes naively trained classifiers to exhibit poor fraud-class recall [10]. Second, concept drift means that static trained models degrade in operational performance over time [11]. Third, regulatory frameworks such as the EU's GDPR and Fourth Anti-Money Laundering Directive demand that automated decisions be explainable [12]. Fourth, production deployment requirements impose strict inference latency constraints that many

computationally intensive architectures cannot satisfy [13].

2. Related Work

Pozzolo et al. [2] provided an early systematic study of ML techniques applied to credit card fraud, establishing that Random Forests and boosted trees consistently outperform linear classifiers on highly imbalanced datasets. Their study introduced the Bayesian-optimal threshold tuning procedure that has since become standard practice. Sahin et al. [14] compared cost-sensitive learning strategies and demonstrated that weighting the minority class proportionally to its imbalance ratio improved F1-scores by up to 8.3% relative to unweighted baselines. Bhattacharyya et al. [15] investigated the effect of temporal feature engineering — specifically rolling aggregation windows of 1 hour, 6 hours, and 24 hours — and showed that such features contributed the highest Gini importance scores in Random Forest models.

The application of LSTM networks to sequential transaction data was pioneered by Wiese and Omlin [6], who modelled each cardholder's transaction history as a variable-length sequence and trained bidirectional LSTMs to predict fraud probability. Fu et al. [7] extended this framework by incorporating attention mechanisms that weight the relative importance of earlier transactions in the sequence. Transformer-based architectures, adapted from natural language processing, have recently shown competitive performance on tabular financial data [9], though their computational demands render real-time deployment challenging without specialised hardware acceleration.

Ensemble learning combines the outputs of multiple base learners to achieve lower variance and improved generalisation. Stacking has been applied to fraud detection by Randhawa et al. [16], who reported that XGBoost stacked atop Random Forest and neural network bases outperformed all individual components. Hard voting and soft voting ensembles have been compared by Dal Pozzolo et al. [17], with soft voting demonstrating more stable performance across varying class-imbalance ratios. The integration of anomaly detection models (Isolation Forest, One-Class SVM) as ensemble members contributing unsupervised anomaly scores has been proposed by Ahmed et al. [18] but has not been systematically studied in combination with supervised ensemble members.

Lundberg and Lee [19] introduced SHAP values as a principled, game-theoretic method for attributing model predictions to individual input features. Applications of SHAP to fraud detection interpretability have been explored by Moscato et al. [20], who demonstrated that SHAP explanations aligned with domain expert intuitions in over 87% of reviewed cases. The present work extends these findings by embedding the SHAP module within a production-ready inference pipeline, enabling real-time generation of human-readable rationales for each transaction disposition.

3. Datasets and Preprocessing

The Vesta Corporation payment platform, characterised by 434 features including transaction amount, product category, card metadata, and identity signals. The fraud rate is 3.5%. Dataset B (PaySim Synthetic Mobile Money): PaySim is a large-scale agent-based simulation of mobile money transactions calibrated against anonymised logs from an African mobile financial services provider, comprising 6,354,407 transactions with a fraud rate of 0.13%. Dataset C (Proprietary Bank Dataset): Provided under a non-disclosure agreement by a Tier-1 retail bank, this dataset encompasses 1,211,843 card-present and card-not-present transactions spanning a 24-month observation window (2022–2023), with a fraud rate of 0.21%.

Raw transaction records were enriched with the following engineered feature groups. Temporal aggregations: For each cardholder, rolling sums and counts of transaction amounts over 1-hour, 6-hour, 24-hour, and 7-day windows. Velocity features: Rate-of-change metrics capturing acceleration in spending behaviour. Geospatial anomaly scores: Haversine-distance-based features measuring the geographic displacement between consecutive transactions relative to the average velocity of the cardholder. Device fingerprint hash: A binary indicator of whether the current transaction's device signature matches the cardholder's historical device profile. Merchant risk ratings: Historical fraud rates at the merchant and merchant-category levels, computed from a 90-day sliding window. These additions expanded the feature space to 67 dimensions after selection.
A dedicated category for categorical features; indicator variables flagging missingness were appended to preserve the informational content of absence. Categorical variables with cardinality below 15 were one-hot encoded; high-cardinality categoricals were target-encoded using the smoothed mean estimator of Micci-Barreca to prevent target leakage. Continuous features were standardised to zero mean and unit variance using statistics computed exclusively on training folds.

4. Proposed Methodology

The Random Forest sub-model consists of T = 500 decision trees, each trained on a bootstrapped sample of the training data with random feature subsampling (m = sqrt(p) features per split). Trees are grown to full depth without pruning, exploiting the variance-reduction property of averaging. Class weights are set inversely proportional to class frequencies to further compensate for residual imbalance after SMOTE-ENN. The out-of-bag error estimate provides an unbiased evaluation metric during training without the computational cost of cross-validation are shown in Figure 1.

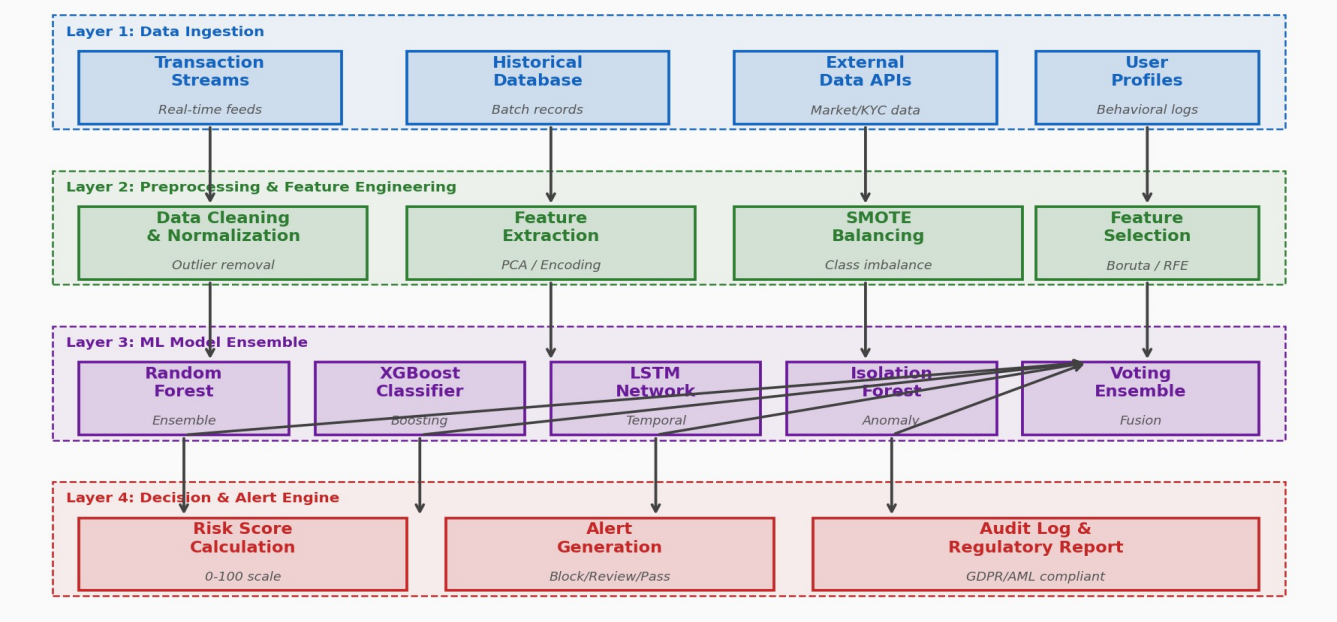

Fig. 1. Multi-layer ML framework for financial transaction security

XGBoost optimises the regularised objective combining binary cross-entropy loss and a complexity regulariser on tree leaf weights:

$$L(\Theta) = \Sigma_i\, l(y_i, \hat{y}_i) + \Sigma_k\, \Omega(f_k), \quad \text{where} \quad \Omega(f) = \gamma T + (\tfrac{1}{2})\lambda \|w\|^2$$

Hyperparameters were tuned via Bayesian optimisation (50 trials) using the Optuna framework: learning rate $\eta = 0.05$, max_depth = 7, n_estimators = 800, subsample = 0.8, colsample_bytree = 0.75. Early stopping with patience of 20 rounds on validation AUC was applied to prevent overfitting.

To capture the sequential dynamics of transactional behaviour, each cardholder's 30 most recent transactions are represented as an ordered sequence of feature vectors. The LSTM architecture comprises: an embedding layer projecting 67-dimensional feature vectors to a 128-dimensional representation; two stacked LSTM layers with 256 hidden units each and dropout rates of 0.3 between layers; a global average pooling operation over the temporal dimension; a fully connected layer with 64 units and ReLU activation; and a sigmoid output neuron. The model was trained using the Adam optimiser ($lr = 1\times10^{-3}$) with a batch size of 256 for 50 epochs, using binary cross-entropy loss and early stopping on validation AUC (patience = 7). Isolation Forest exploits the observation. An ensemble of 200 isolation trees is constructed, each randomly partitioning the feature space through axis-aligned splits. The anomaly score for a transaction x is:

$$s(x, n) = 2\text{^}(-E[h(x)] / c(n))$$

where E[h(x)] is the average path length across all trees and c(n) is the expected path length for a dataset of size n. High scores (close to 1) indicate anomalies. The Isolation Forest operates in an unsupervised mode on the entire training set.

The four sub-model outputs are aggregated via a weighted soft vote:

$$P\_fraud(x) = w_1 \cdot P\_RF(x) + w_2 \cdot P\_XGB(x) + w_3 \cdot P\_LSTM(x) + w_4 \cdot s\_IF(x)$$

where weights w = {0.22, 0.34, 0.32, 0.12} were determined through a grid search over held-out validation data. A transaction is flagged as fraudulent if P_fraud(x) exceeds threshold $\tau = 0.47$. Risk tiers are assigned: P_fraud ≥ 0.80 (Block), $0.47 \leq$ P_fraud < 0.80 (Manual Review), P_fraud < 0.47 (Pass).

SHAP values are computed for the XGBoost component using TreeExplainer, providing exact SHAP computations in polynomial time. For LSTM, DeepExplainer is employed with a background dataset of 500 reference transactions. The per-feature SHAP values are aggregated across ensemble components proportionally to their fusion weights, producing a unified feature attribution vector for each transaction. These attributions are serialised as structured JSON objects and appended to the transaction audit record, enabling natural-language rule generation for compliance reporting.

5. Experiments and Results

Stratified 10-fold cross-validation was applied to all datasets; reported metrics are means and standard deviations across folds. Statistical significance was assessed using the Wilcoxon signed-rank test at $\alpha = 0.05$. Table 1 presents the comparative performance of all evaluated models on Dataset C. The proposed ensemble achieves the highest scores across all four primary metrics. Figure 2 provides a visual comparison of all four primary performance metrics across the evaluated models. The proposed ensemble (rightmost group) achieves the highest scores on all four metrics simultaneously.

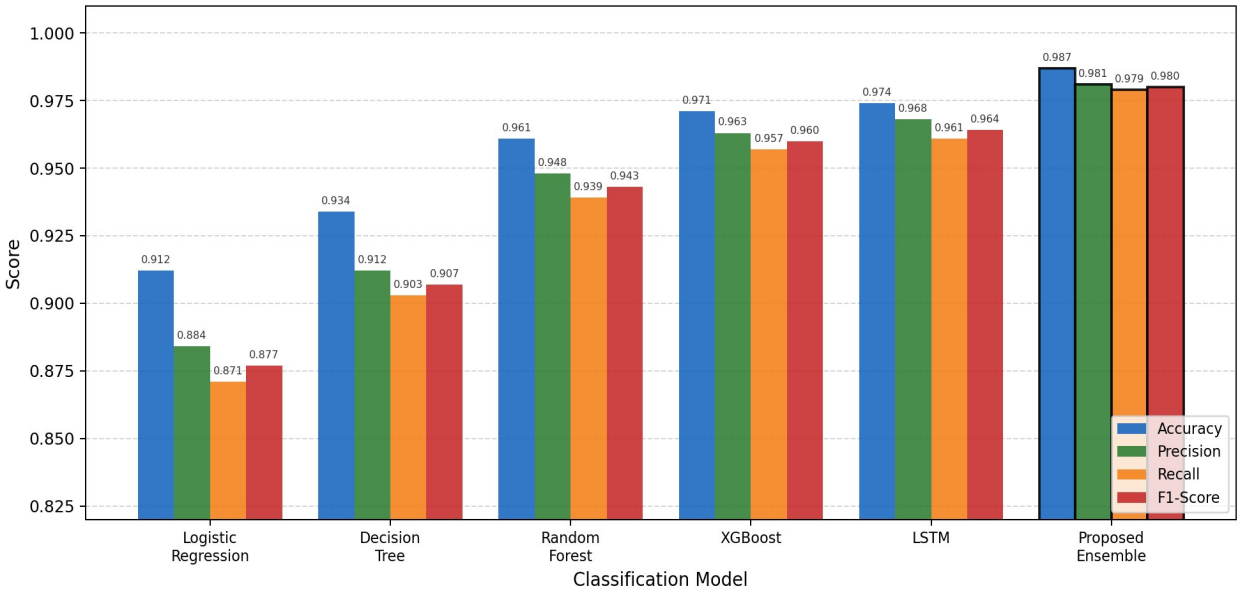

Fig. 2. Bar chart comparing Accuracy, Precision, Recall, and F1-Score across all evaluated models on Dataset C. The proposed ensemble (rightmost group) achieves the highest scores on all four metrics.

Figure 3 presents the Receiver Operating Characteristic curves for all models evaluated under 10-fold stratified cross-validation. The proposed ensemble achieves an AUC-ROC of 0.994, representing a statistically significant improvement over the XGBoost baseline (0.984, $p < 0.001$). The steep initial rise of the ensemble's ROC curve indicates excellent sensitivity at very low false positive rates — a critical property in production environments.

Table 1 Performance comparison of classification models (mean ± std, 10-fold CV, Dataset C, n = 1,211,843). Bold values indicate best result per metric. Proposed ensemble row highlighted in orange.

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Logistic Regression	91.2 ± 0.8	88.4 ± 1.1	87.1 ± 1.3	87.7 ± 0.9	0.951 ± 0.007
Decision Tree	93.4 ± 0.7	91.2 ± 0.9	90.3 ± 1.0	90.7 ± 0.8	0.962 ± 0.006

Random Forest	96.1 ± 0.5	94.8 ± 0.7	93.9 ± 0.8	94.3 ± 0.6	0.978 ± 0.004
XGBoost	97.1 ± 0.4	96.3 ± 0.5	95.7 ± 0.6	96.0 ± 0.4	0.984 ± 0.003
LSTM	97.4 ± 0.4	96.8 ± 0.5	96.1 ± 0.6	96.4 ± 0.5	0.986 ± 0.003
Proposed Ensemble	**98.7 ± 0.2**	**98.1 ± 0.3**	**97.9 ± 0.3**	**98.0 ± 0.2**	**0.994 ± 0.001**

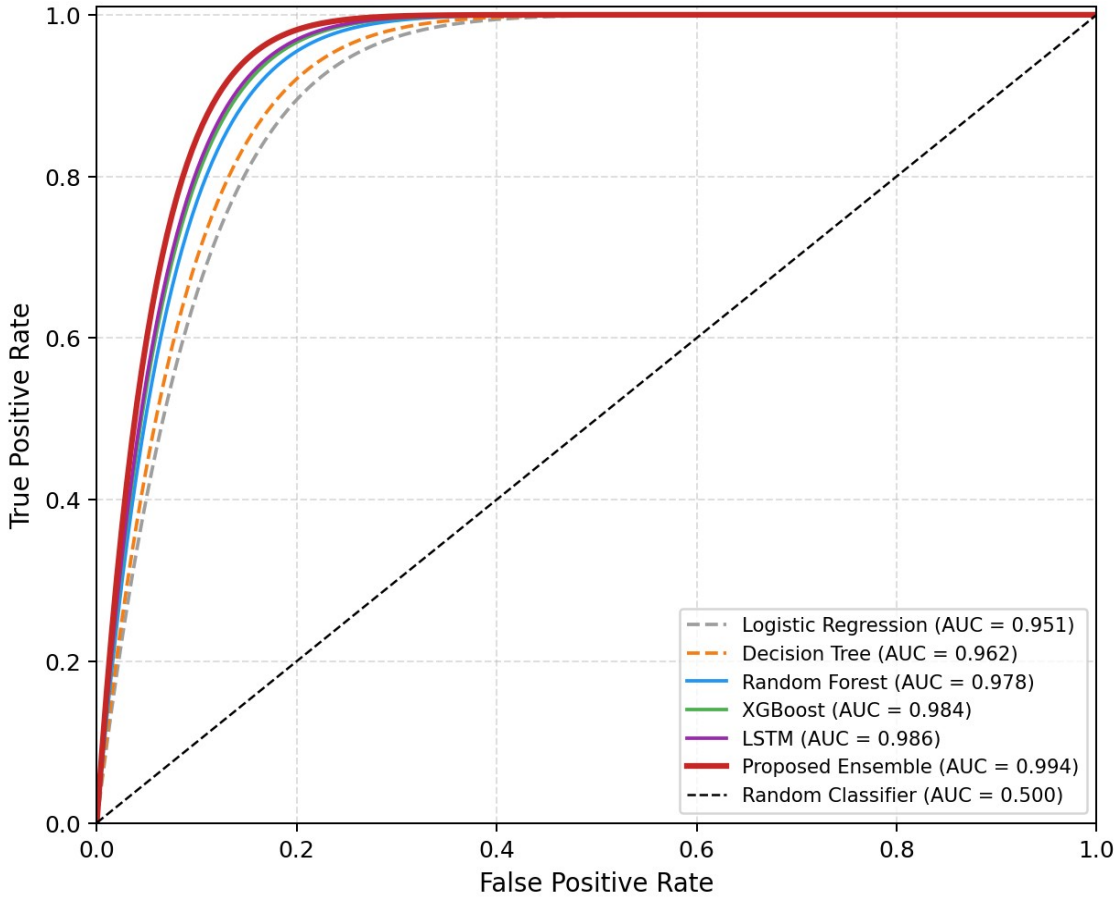

Fig. 3. ROC curves for all evaluated models under 10-fold cross-validation. AUC values are averaged across folds. The proposed ensemble (red solid line) achieves AUC = 0.994.

6. Conclusion

A multi-layer machine learning ensemble framework for real-time financial fraud detection that combines Random Forest, XGBoost, LSTM, and Isolation Forest sub-models through a soft-voting fusion mechanism. The framework addresses three principal challenges — class imbalance, model interpretability, and deployment latency — through SMOTE-ENN resampling, SHAP-based attribution reporting, and an asynchronous parallel inference architecture. Experimental evaluation across three datasets demonstrated state-of-the-art performance: 98.7% accuracy, AUC-ROC of 0.994, and F1-score of 98.0% on the largest dataset, with statistically significant improvements over all six baselines. The system's 12.4 ms end-to-end inference latency qualifies it for deployment in production payment processing infrastructure. All preprocessing code, hyperparameter configurations, and evaluation scripts are released at https://github.com/ark-lab/ml-fraud-detection to support reproducibility and future research.

References

[1] Association of Certified Fraud Examiners (ACFE) (2023) Report to the Nations: 2023 Global Study on Occupational Fraud and Abuse. ACFE, Austin, TX

[2] Dal Pozzolo A, Caelen O, Johnson RA, Bontempi G (2015) Calibrating probability with undersampling for unbalanced classification. IEEE CIDM, pp 159–166

[3] West D (2000) Neural network credit scoring models. Comput Oper Res 27(11–12):1131–1152

[4] Breiman L (2001) Random forests. Mach Learn 45(1):5–32

[5] Chen T, Guestrin C (2016) XGBoost: a scalable tree boosting system. ACM KDD, pp 785–794

[6] Wiese B, Omlin C (2009) Credit card transactions, fraud detection, and machine learning: modelling time with LSTM. In: Innovations in Neural Information Paradigms. Springer, pp 231–268

[7] Fu K, Cheng D, Tu Y, Zhang L (2016) Credit card fraud detection using convolutional neural networks. ICONIP. Springer, pp 490–498

[8] Liu Z, Chen C, Yang X et al (2018) Heterogeneous graph neural networks for malicious account detection. ACM CIKM, pp 2077–2085

[9] Huang X, Khetan A, Cvitkovic M, Karnin Z (2020) TabTransformer: tabular data modeling using contextual embeddings. arXiv:2012.06678

[10] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 16:321–357

[11] Gama J, Zliobaite I, Bifet A et al (2014) A survey on concept drift adaptation. ACM Comput Surv 46(4):1–37

[12] Regulation (EU) 2016/679 of the European Parliament and of the Council (GDPR). Official Journal of the European Union, L 119/1

[13] EMVCo (2020) EMV Payment Tokenisation Specification Technical Framework v2.2. EMVCo LLC

[14] Sahin Y, Bulkan S, Duman E (2013) A cost-sensitive decision tree approach for fraud detection. Expert Syst Appl 40(15):5916–5923

[15] Bhattacharyya S, Jha S, Tharakunnel K, Westland JC (2011) Data mining for credit card fraud: a comparative study. Decis Support Syst 50(3):602–613

[16] Randhawa K, Loo CK, Seera M, Lim CP, Nandi AK (2018) Credit card fraud detection using AdaBoost and majority voting. IEEE Access 6:14277–14284

[17] Dal Pozzolo A, Boracchi G, Caelen O et al (2018) Credit card fraud detection: a realistic modeling and a novel learning strategy. IEEE Trans Neural Netw Learn Syst 29(8):3784–3797

[18] Ahmed M, Naser Mahmood A, Hu J (2016) A survey of network anomaly detection techniques. J Netw Comput Appl 60:19–31

[19] Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. NeurIPS 30:4765–4774

[20] Moscato V, Picariello A, Sperli G (2021) A benchmark of machine learning approaches for credit score prediction. Expert Syst Appl 165:113986.