

A Layered Cybersecurity Framework for Data Protection and Privacy-Preserving Fraud Detection in Digital Financial Transactions

Abhishek Ramprasad Kharde, Prof. Bangar Abhishek Popat,
Prof. Bhosale Sachin Bajirao, Dr. A. A. Khatri

Department of Computer Engineering, Jaihind College of Engineering, Pune, Maharashtra, India

Abstract — Digital payments have made everyday life easier, but they have also given attackers a larger and more profitable surface to work on. Banks and payment processors now face two problems at the same time: stopping fraudulent transactions quickly, and protecting the personal data that fraud models need in order to learn. These two goals often pull in opposite directions, because stronger detection usually means collecting and sharing more sensitive data. This paper proposes CPDF, a five-layer framework that treats data protection and fraud detection as one design problem instead of two separate ones. The layers cover secure ingestion, privacy preservation through pseudonymisation, minimisation and differential privacy, hybrid detection that pairs an unsupervised Isolation Forest outlier score with a class-weighted gradient-boosted classifier, explainable alerts using SHAP, and continuous monitoring aligned with the NIST Cybersecurity Framework 2.0. We describe the architecture, the data flow, the evaluation protocol and the threats the design is meant to withstand. The work is a proposed design: it reports no experimental results, and the expected outcomes are stated as testable hypotheses for future implementation.

Keywords— cybersecurity, data protection, fraud detection, differential privacy, explainable machine learning, federated learning.

I. INTRODUCTION

1. Background

Fraud has followed money for as long as money has existed, but the move to cards, wallets and instant bank transfers changed its scale and its speed. A stolen card number can now be tested within seconds from anywhere in the world. Researchers have studied statistical and data-mining approaches to fraud for more than two decades [1], [2], and the field has grown from simple rule lists to ensembles and deep models. Meanwhile, the data these systems depend on has become more tightly regulated. The European General Data Protection Regulation sets out principles such as data minimisation, purpose limitation and data protection by design [3], and India's Digital Personal Data Protection Act, 2023 places similar duties of consent and care on organisations that handle personal data [4]. A fraud team in 2026 therefore cannot simply gather everything and sort it out later.

2. Problem Statement

Most fraud-detection research optimises a single quantity, usually detection accuracy, and treats the security of the data pipeline as someone else's job. In practice that split does not hold. A model trained on raw cardholder data becomes a sensitive asset in its own right, and a breach of the training store can be as damaging as the fraud it was meant to prevent. What

is missing is a design that keeps the detection pipeline, the privacy controls and the audit trail in one coherent structure.

3. Need for the Study

There are three reasons this matters now. First, fraud is extremely rare compared with legitimate activity, so models can look accurate while missing nearly every fraudulent case.

Second, institutions are reluctant to pool customer data across organisations, even though fraud rings do not respect institutional borders. Third, analysts and regulators increasingly ask why a transaction was blocked, and a score without a reason is hard to defend.

4. Objectives

The objectives of this work are: (1) to define a layered architecture that embeds data protection controls inside a fraud-detection pipeline; (2) to specify a hybrid detector suited to highly imbalanced data; (3) to describe where differential privacy and federated learning can be applied without breaking usability; (4) to provide explanations that an analyst can act on; and (5) to set out an evaluation protocol that measures detection quality, privacy loss and operational cost together.

II. LITERATURE REVIEW

Early surveys laid the groundwork. Bolton and Hand reviewed statistical fraud detection and stressed that fraud is an adversarial, shifting problem [1]. Ngai et al. classified data-mining applications in financial fraud and showed how heavily the literature leaned on classification and clustering [2]. Abdallah et al. later surveyed fraud detection systems and highlighted concept drift, class imbalance and the lack of public data as persistent obstacles [5].

On realistic modelling, Dal Pozzolo et al. argued that fraud detection should be judged the way analysts actually work, where only a small number of alerts can be reviewed each day and labels arrive late [6]. Carcillo et al. combined unsupervised outlier scores with supervised learning on real card data and reported that the combination helped the supervised models [7]. That finding motivates the hybrid layer we propose. The unsupervised half draws on Isolation Forest, which isolates anomalies through random partitioning and scales well to large data [8], while the supervised half draws on gradient boosting as implemented in XGBoost [9].

Class imbalance has its own line of work. Chawla et al. introduced SMOTE, which creates synthetic minority samples [10]. Saito and Rehmsmeier showed that precision-recall plots give a more honest picture than ROC curves when positives are rare [11], which is why our protocol relies on precision-recall measures.

For privacy, Dwork and Roth set out the formal foundations of differential privacy [12]. Abadi et al. showed how to train deep networks with a differentially private version of stochastic gradient descent and track the privacy budget [13]. McMahan et al. introduced federated averaging so that devices or institutions can train a shared model without exchanging raw data [14], and Bonawitz et al. described a secure aggregation protocol so that the server sees only the combined update [15]. Federated learning is not a complete shield, however: Zhu et al. demonstrated that shared gradients can leak training samples [16], and Kairouz et al. catalogued open problems in the area, including privacy, robustness and communication cost [17]. Yang et al. gave a useful taxonomy of horizontal, vertical and transfer settings [18], and a cross-bank fraud scenario falls naturally into the horizontal case.

For explanation, Lundberg and Lee unified several attribution methods under SHAP [19], and Ribeiro et al. proposed LIME for local surrogate explanations [20]. Both are widely used to justify individual predictions. Finally, the NIST Cybersecurity

Framework 2.0 organises security work into six functions: Govern, Identify, Protect, Detect, Respond and Recover [21]. We use it as a governance reference for the monitoring layer.

Table 1. Summary of related work and the gap it leaves

Work	Contribution	Gap for our setting
[6] Dal Pozzolo et al.	Realistic fraud evaluation with delayed labels	No privacy layer
[7] Carcillo et al.	Unsupervised plus supervised hybrid	Raw data assumed shareable
[13] Abadi et al.	Differentially private deep learning	Not fraud-specific
[14], [15] McMahan; Bonawitz	Federated training; secure aggregation	Gradient leakage remains [16]
[19] Lundberg and Lee	SHAP explanations	Needs privacy-aware use
[21] NIST	Security governance functions	Not tied to ML pipelines

Limitations of Existing Techniques

Looking across this literature, a handful of weaknesses keep recurring. Rule-based systems are easy to audit but are quickly learned and bypassed by determined fraudsters. Purely supervised models depend on labels that are scarce, delayed and sometimes wrong. Purely unsupervised detectors raise many false alarms because unusual does not always mean fraudulent. Oversampling methods such as SMOTE can blur class boundaries if they are applied before the data is split. Privacy techniques, for their part, cost something: differential privacy adds noise that reduces accuracy, and federated learning adds communication overhead and still exposes gradient information [16]. Finally, few studies measure detection, privacy and explanation within the same experiment, so trade-offs stay hidden.

III. PROPOSED METHODOLOGY

1. Data and Preparation

We plan to evaluate the framework on the widely used credit card fraud dataset released by the Machine Learning Group of Université Libre de Bruxelles in collaboration with Worldline. It holds 284,807 transactions made by European cardholders over two days in September 2013, of which only 492 are

fraudulent, about 0.172 percent [6]. Its features are principal components of the original variables, plus time and amount, which also means it carries no direct identifiers. Fig. 3 shows how lopsided the classes are.

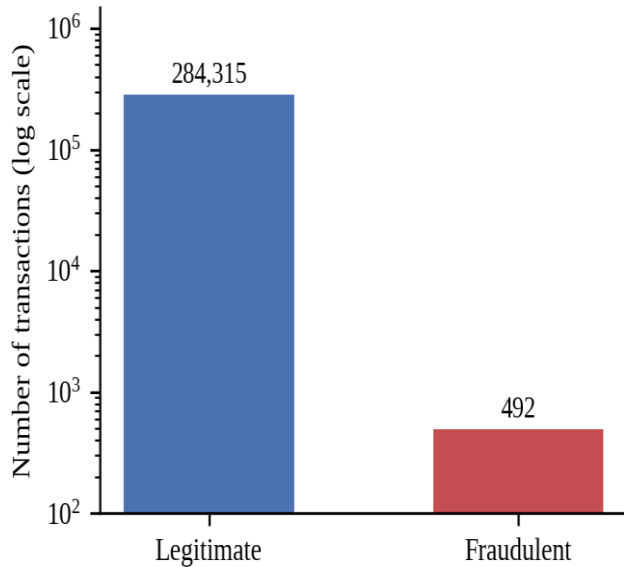


Fig. 3. Class distribution in the ULB credit card fraud dataset (284,315 legitimate and 492 fraudulent transactions, log scale). Because fraud patterns change with time, the data will be split chronologically, with earlier transactions for training and later ones for testing, rather than shuffled at random. Any resampling or class weighting will be fitted inside the training portion only, so that no information flows from the test period into the model.

2. Layer 1 and Layer 2: Protection and Privacy

The first layer secures the pipeline itself. Data in transit uses TLS 1.3, data at rest uses AES-256 in an authenticated mode, direct identifiers such as card numbers are replaced by tokens, and access follows role-based least-privilege rules. The second layer reduces what the models can learn about any individual. It applies minimisation by dropping fields that the detector does not need, pseudonymisation of the remaining identifiers, and differential privacy for statistics or model updates that leave the institution. For a numeric query with sensitivity Δf , the Laplace mechanism adds noise drawn from a Laplace distribution with scale $b = \Delta f / \epsilon$ [12]. Smaller values of ϵ give stronger privacy but more noise, as Fig. 4 illustrates.

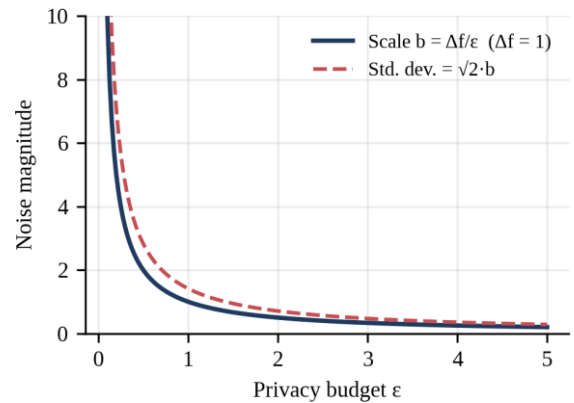


Fig. 4. Laplace noise magnitude against privacy budget ϵ for unit sensitivity (analytical relationship, not experimental data). When several banks take part, each one trains locally and shares only model updates. Updates are clipped, perturbed in the manner of DP-SGD [13] and combined through secure aggregation [15], which limits the gradient-leakage risk described in [16].

3. Layer 3: Hybrid Detection

The detector has two stages. An Isolation Forest [8] is trained on the training data and produces an anomaly score for each transaction. That score is appended to the feature vector, and a gradient-boosted classifier [9] is trained on the enriched features, with class weights set inversely to class frequency. Class weighting is preferred as the default because it does not create synthetic records; SMOTE [10] is kept as a comparison baseline. The decision threshold is not fixed at 0.5. It will be tuned on a validation period to match the number of alerts analysts can realistically review, in line with the argument in [6].

4. Layers 4 and 5: Explanation and Monitoring

For every alert the system attaches SHAP values [19] showing which features pushed the score up, so an analyst sees a short list of reasons instead of a bare number. Analyst decisions are fed back as new labels. The fifth layer keeps tamper-evident logs, watches input and score distributions for drift, and maps controls to the NIST functions [21]. This makes the framework auditable, which matters under regulations such as [3] and [4].

IV. SYSTEM ARCHITECTURE AND FLOWCHART

Fig. 1 shows the five layers and Fig. 2 shows the path that a single transaction takes through the system. A transaction is

first validated, encrypted and tokenised, then enriched with time-aware features such as recent transaction counts and amounts. Privacy controls are applied before the data reaches the models. The Isolation Forest score and the classifier score are combined into a risk value. If that value exceeds the threshold, an alert with explanations goes to the analyst; otherwise the transaction proceeds and is logged.

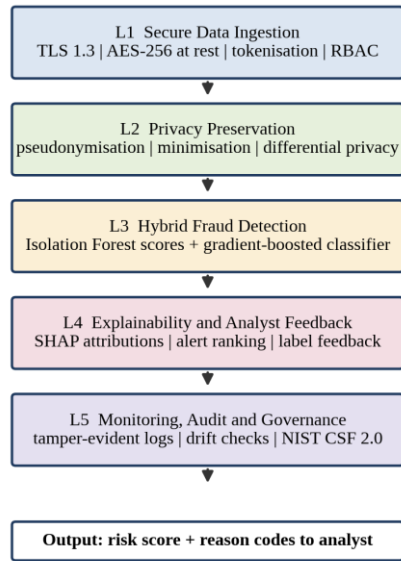


Fig. 1. Five-layer architecture of the proposed CPDF framework.

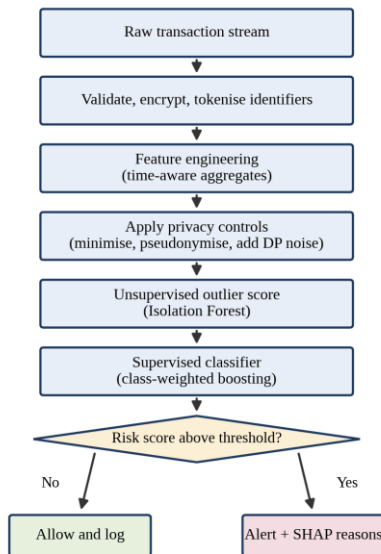


Fig. 2. Flowchart of transaction processing in CPDF.

V. EVALUATION PLAN AND EXPECTED OUTCOMES

Since this paper proposes a design, we do not report measured results. Instead we state what we expect to test, and a failure to confirm any of these expectations would be an informative finding in itself. Table II lists the planned measures.

Table 2. Planned evaluation measures

Aspect	Measure
Detection quality	Area under the precision-recall curve; recall at a fixed alert budget; F1; Matthews correlation coefficient
Privacy	Privacy budget ϵ used; effect of ϵ on detection quality
Explainability	Stability of top SHAP features across runs; analyst review
Operations	Scoring latency per transaction; training time; communication volume in federated runs

We expect that (H1) the hybrid detector will beat the supervised-only and unsupervised-only baselines on precision-recall measures, following the direction reported in [7]; (H2) moderate privacy budgets will cost a small, measurable drop in detection quality, while very strict budgets will cost noticeably more, in line with the relationship in Fig. 4; (H3) class weighting will perform at least as well as SMOTE without the need for synthetic records; and (H4) SHAP explanations will remain stable enough to support analyst decisions. These are hypotheses, not findings.

VI. CONCLUSION

Fraud detection and data protection are usually discussed by different communities, yet in a real bank they share the same data, the same servers and the same regulators. We proposed CPDF, a layered framework that places encryption, privacy controls, hybrid detection, explanation and auditing inside one pipeline, and we laid out a protocol to test it honestly on imbalanced, time-ordered data. The next step is implementation and measurement, and we are careful not to claim performance that has not been demonstrated.

Future Scope

Several extensions look promising. Graph-based models could capture links between accounts and merchants. Sequence models could learn customer behaviour over time. Stronger

privacy tools such as homomorphic encryption or trusted execution environments could be compared with differential privacy. Adversarial testing, where an attacker deliberately probes the model, should also be added, as should a study with real analysts to check that the explanations truly help them. Finally, testing on additional public and, where permitted, institutional datasets would show how well the findings generalise.

REFERENCES

1. R. J. Bolton and D. J. Hand, "Statistical fraud detection: A review," *Statistical Science*, vol. 17, no. 3, pp. 235–255, 2002.
2. E. W. T. Ngai, Y. Hu, Y. H. Wong, Y. Chen, and X. Sun, "The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature," *Decision Support Systems*, vol. 50, no. 3, pp. 559–569, 2011.
3. Regulation (EU) 2016/679 of the European Parliament and of the Council (General Data Protection Regulation), *Official Journal of the European Union*, L 119, 2016.
4. Government of India, *The Digital Personal Data Protection Act*, 2023, 2023.
5. A. Abdallah, M. A. Maarof, and A. Zainal, "Fraud detection system: A survey," *Journal of Network and Computer Applications*, vol. 68, pp. 90–113, 2016.
6. A. Dal Pozzolo, G. Boracchi, O. Caelen, C. Alippi, and G. Bontempi, "Credit card fraud detection: A realistic modeling and a novel learning strategy," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 8, pp. 3784–3797, 2018.
7. F. Carcillo, Y.-A. Le Borgne, O. Caelen, Y. Kessaci, F. Oblé, and G. Bontempi, "Combining unsupervised and supervised learning in credit card fraud detection," *Information Sciences*, vol. 557, pp. 317–331, 2021.
8. F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, 2008, pp. 413–422.
9. T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016, pp. 785–794.
10. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.
11. T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," *PLoS ONE*, vol. 10, no. 3, e0118432, 2015.
12. C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Foundations and Trends in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–407, 2014.
13. M. Abadi et al., "Deep learning with differential privacy," in *Proc. ACM SIGSAC Conf. Computer and Communications Security (CCS)*, 2016, pp. 308–318.
14. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. 20th Int. Conf. Artificial Intelligence and Statistics (AISTATS)*, PMLR vol. 54, 2017, pp. 1273–1282.
15. K. Bonawitz et al., "Practical secure aggregation for privacy-preserving machine learning," in *Proc. ACM SIGSAC Conf. Computer and Communications Security (CCS)*, 2017, pp. 1175–1191.
16. L. Zhu, Z. Liu, and S. Han, "Deep leakage from gradients," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, 2019.
17. P. Kairouz et al., "Advances and open problems in federated learning," *Foundations and Trends in Machine Learning*, vol. 14, no. 1–2, pp. 1–210, 2021.
18. Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," *ACM Trans. Intell. Syst. Technol.*, vol. 10, no. 2, Art. 12, 2019.
19. S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems (NIPS)*, vol. 30, 2017, pp. 4765–4774.
20. M. T. Ribeiro, S. Singh, and C. Guestrin, "“Why should I trust you?” Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144.
21. National Institute of Standards and Technology, *The NIST Cybersecurity Framework (CSF) 2.0*, NIST CSWP 29, 2024.