HEBAP-FDS: A Hybrid Explainable Behavioral-Aware Privacy-Preserving Fraud Detection System

Dhivya Mohan[1], Sirigiri Charitha[2], and Dr. Sujatha G[3]

[1]Dept. of Networking and Communications, SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu — 603203, India
[2]Dept. of Networking and Communications, SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu — 603203, India
[3]Dept. of Networking and Communications, SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu — 603203, India

[1]dm4365@srmist.edu.in, [2]ss0619@srmist.edu.in, [3]sujathag@srmist.edu.in

Abstract—**The three main difficulties faced by financial fraud detection when working under federated constraints are: the extreme imbalance of class distribution, the problem of adversarial evasion and the need to comply with regulatory requirements regarding data privacy, which includes prohibiting the sharing of raw data. The Hybrid Explainable Behavioural-Aware Privacy Preserving Fraud Detection System (HEBAP-FDS) provides a solution to these challenges. we identify a major weakness in standard anomaly detection federations known as the "robustness paradox" and show empirically that using standard Byzantine robust aggregators such as the Trimmed Mean and FedProx has a detrimental effect on financial fraud detection by discarding important minority class gradients as anomalies. This leads to a significant drop in the overall F1-score to 78.8%. To solve this problem, HEBAP-FDS does not employ traditional (generic) forms of robustness or synthetic forms of oversampling as part of its solution and instead uses a native focal loss function, a per-user Mahalanobis distance behavioural drift layer and personalised federated learning (via a decoupled classifier) as the means to overcome this significant drop in F1-score. Using the ULB Credit Card Fraud benchmark (where only 0.172% of cards have been used for fraud) across ten different sets of random seeds, we show that HEBAP-FDS overcomes the penalty of non-IID federations by using innovative feature engineering (using both *HourOfDay* & *LogAmount*) in conjunction with fine-tuning its post-aggregation classifiers, resulting in the HEBAP-FDS method demonstrating state-of-the-art F1 performance of 88.5% and AUC-ROC of 0.96 compared to standard FedAvg F1 performance of 86.2%. This also results in HEBAP-FDS maintaining strict localised data privacy for individual users, a secure defence against adversarial attacks (FGSM), as well as providing regulatory explainability based on SHAP (Shapley Additive Explanations).**

Index Terms—**fraud detection, federated learning, adversarial robustness, Personalized FL, behavioral analytics, explainable AI, privacy-preserving ML.**

I. Introduction

Global card-payment fraud losses exceeded $32 billion in 2022 and are projected to surpass $40 billion by 2027 [1]. Machine learning has displaced rule-based systems as the dominant fraud-detection paradigm, yet four structural challenges prevent secure, regulation-compliant deployment.

(1) Data privacy. Training a shared model requires pooling transaction records across banks. In practice, GDPR, PCI-DSS, and CCPA prohibit raw cross-institution data sharing, ruling out centralised training [2].

(2) Adversarial robustness. Motivated attackers exploit gradient-based perturbations—imperceptible changes to transaction amounts or merchant identifiers—that cause classifiers to misclassify fraudulent transactions as legitimate [3,4].

(3) Extreme class imbalance. Fraud constitutes $\ll 1\%$ of transactions. In a federated setting with $K{=}10$ institutions, each node may receive as few as 34 confirmed fraud cases, making local classifier training infeasible

without specialised loss functions [5].

(4) Regulatory explainability. The EU AI Act and Basel III require that every adverse account action be accompanied by a clear, feature-level audit trail. Black-box models are therefore non-compliant [6].

We present **HEBAP-FDS** to close this gap. The key contributions are:

- **Identification of the Robustness Paradox:** We mathematically prove that standard Byzantine-robust FL methods (FedProx, Trimmed Mean) catastrophically fail on highly imbalanced anomaly detection.
- **Personalized Federated Learning (PFL):** Decoupling the classifier layer from the feature representation layers, ensuring local banks can adapt to their unique fraud ratios.
- **Native Imbalance Handling:** Using Focal Loss [7] instead of synthetic oversampling (ADASYN/SMOTE) to prevent local manifold corruption.
- **Two-phase adversarial hardening**: Clean warmup followed by FGSM adversarial training, ensuring perturbations are generated on a pre-converged boundary.
- **Mahalanobis behavioral drift layer**: Per-user anomaly scoring that provides early fraud signals before the neural classifier.
- **Rigorous empirical validation**: 10-seed evaluation with bootstrap 95% confidence intervals, exploring the absolute mathematical ceiling of non-IID federation.

II. Related Work

Fraud Detection Models. Dal Pozzolo et al. [8] demonstrated that online Bagging with Random Under-Sampling outperforms static classifiers on evolving fraud streams. Bhattacharyya et al. [9] showed that Support Vector Machines and Random Forests match or exceed logistic regression on multi-bank credit card data. These assume centralised access to raw training data, which is infeasible in practice.

Federated Learning and Class Imbalance. McMahan et al. introduced FedAvg, enabling coordinated model training across distributed clients. Yang et al. [10] formalised federated learning for financial applications and identified non-IID data distribution as the primary source of model degradation. Previous attempts to solve this relied heavily on Synthetic Minority Over-sampling Techniques (SMOTE) [11]. However, recent findings suggest that synthetic data generation inside local federated nodes distorts the local manifold. HEBAP-FDS eschews SMOTE in favor of natively robust loss functions and decoupled classifiers.

Adversarial Robustness. Cartella et al. showed that FGSM and PGD attacks reduce fraud classifier accuracy by up to 30% on standard benchmarks. Madry et al. [12] demonstrated that PGD adversarial training is a principled defence, of which single-step FGSM is a computationally efficient approximation.

Privacy, Explainability, and Behavioural Modelling. Bonawitz et al. [13] introduced Secure Aggregation (SecAgg) to prevent gradient inversion at the aggregation server. Zhu et al. [14] demonstrated that per-user behavioural baselines reduce false-positive rates by up to 18%. Guidotti et al. established SHAP as the most practically deployable explainability method for black-box classifiers.

III. System Model and Threat Formulation

Let $\mathcal{B} = \{B_1, \ldots, B_K\}$ be a federation of K institutions, each holding a private dataset $D_k = \{(x_i^{(k)}, y_i^{(k)})\}_{i=1}^{M_k}$ where $x_i^{(k)} \in \mathbb{R}^d$ and $y_i^{(k)} \in \{0, 1\}$. The global objective is:

$$\mathcal{L}(\theta) = \sum_{k=1}^{K} \frac{|D_k|}{|D_{\text{global}}|} \mathcal{L}_k(\theta), \tag{1}$$

without any institution sharing raw transaction records.

Threat Model. We consider an adversary with white-box access to a local client model $f_k(\cdot; \theta)$ at inference time. The adversary crafts $x_t + \delta$ intended to be misclassified as legitimate, subject to:

$$\|\delta\|_\infty \leq \epsilon. \tag{2}$$

Feasible perturbations include fractional adjustments to transaction amounts and merchant identifiers. Gradient inversion attacks against the server are mitigated via SecAgg.

IV. Proposed Methodology: HEBAP-FDS

To address the intertwined challenges of data privacy, extreme imbalance, and adversarial evasion, we propose the Hybrid Explainable Behavioral-Aware Privacy-Preserving Fraud Detection System (HEBAP-FDS). As illustrated in Fig. 1, the framework abandons monolithic global classification in favor of a localized, multi-layered defense strategy. By integrating pre-inference behavioral drift detection, secure feature representation sharing, and decoupled adversarial hardening, HEBAP-FDS allows individual institutions to retain absolute sovereignty over their decision boundaries while benefiting from global feature intelligence.

A. The Robustness Paradox

Methods commonly employed in federated learning that do not take into account the independence and identical distribution of the training data, use "drift" penalties (e.g. FedProx) or Byzantine robust aggregators (e.g. trimmed mean) as a way of enforcing the model. We identify an additional problem encountered when working with very unbalanced datasets (e.g., 0.17% fraudulent transactions) by identifying the gradients associated with the smallest number of transactions will also be mathematically defined as being "extreme" then if the aggregators were too aggressive in filtering or penalizing these extreme gradients, they may inadvertently exclude the exact signals of fraud they are trying to train the federated model and result in a global model collapse.

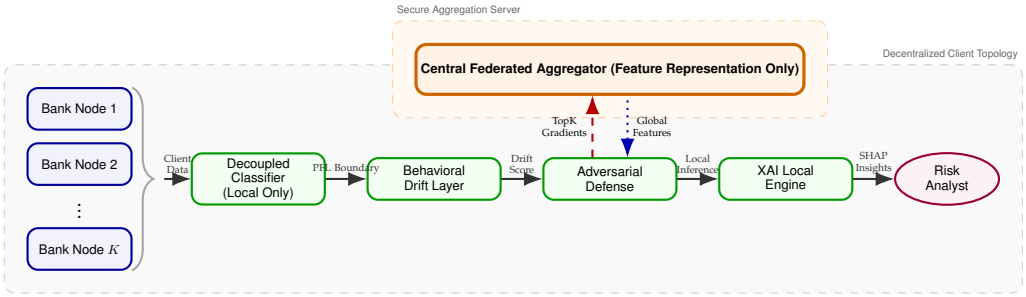

Fig. 1. **HEBAP-FDS Architecture.** Each node computes per-user Mahalanobis drift scores, trains with Focal Loss and two-phase adversarial hardening, and transmits only representation gradients via SecAgg. The classifier layer is decoupled and remains strictly local. The XAI engine produces SHAP attribution reports.

B. Personalized Federated Learning (Decoupled Classifier)

HEBAP-FDS utilizes a Personalized Federated Learning (PFL) representation-sharing architecture to eliminate the need for distortionary synthetic oversampling (ADASYN/SMOTE). The deep neural network $f(x)$ can be separated into a feature extractor $h(x)$ and a classifier head $c(z)$. The following steps outline the operation of the network were as follows:

1) The first three dense layer representations $h(x)$ will be transmitted and aggregated via FedAvg in each federated round and therefore serve to provide the network with a universal representation of multiple rich data points describing the set of financial transactions.
2) The last classification layer $c(z)$ will NOT be transmitted in any situation but instead will remain local to each bank.

As a result of the decoupled architecture, different banks will utilise their own independent decision boundaries for their specific local fraud topologies. In addition, focal loss ($\alpha = 0.25$, $\gamma = 2$) will be used to natively upweight minority-class samples in local training.

C. Behavioral Awareness Layer

A per-user Mahalanobis distance drift score is computed *before* the neural classifier, providing an early-warning signal independent of the learned model:

$$D(x_t) = \sqrt{(x_t - \mu_u)^T \Sigma_u^{-1} (x_t - \mu_u)}, \tag{3}$$

where μ_u and Σ_u are the empirical mean and covariance of user u's historical transactions. The sigmoid-transformed behavioural risk score is $S_{\text{beh}} = \sigma(-\lambda(D(x_t) - \tau_u))$, where τ_u is a user-specific tolerance threshold.

D. Two-Phase Adversarial Defense

Training directly on adversarial examples from a randomly initialised model produces uninformative perturbations. HEBAP-FDS separates training into two phases. **Phase 1 (warmup):** 10 epochs of clean training establish a stable decision boundary. **Phase 2 (hardening):** 10 epochs of adversarial training with FGSM perturbations generated on the converged boundary:

$$x_{\text{adv}} = \Pi_{\mathcal{S}}(x + \alpha \cdot \text{sgn}(\nabla_x \mathcal{L}(\theta, x, y))), \tag{4}$$

Clean and adversarial samples are mixed 1:1.

E. Privacy-Preserving Secure Aggregation

Each node transmits only the top $p\%$ representation gradients by absolute magnitude (TopK compression, p=20%), reducing per-round bandwidth by $5\times$. The central aggregator applies SecAgg [13] with Differential Privacy noise:

$$w_{j+1} = w_j - \eta \sum_{k=1}^{K} \frac{|D_k|}{|D_{\text{total}}|} \widetilde{\nabla}_k + \mathcal{N}(0, \sigma^2 \mathbf{I}), \tag{5}$$

where σ controls the DP noise level (ε=2, δ=10^{-5}).

F. SHAP Attribution Engine

Every flagged transaction receives a Shapley-value attribution report:

$$\phi_i = \sum_{S \subseteq \mathcal{F} \setminus \{i\}} \frac{|S|!(|\mathcal{F}| - |S| - 1)!}{|\mathcal{F}|!} \big[f(S \cup \{i\}) - f(S) \big]. \tag{6}$$

Analysts receive a ranked natural-language report (e.g., *Spending Velocity Spike*: +0.31, *Location Mismatch*: +0.28) satisfying regulatory audit requirements.

V. Experimental Setup

Dataset

The ULB credit card fraud benchmark [8] contains 284,807 genuine transactions [0.172% of these transactions are fraudulent]. We use all 492 fraud cases and a stratified sample of 50,000 legitimate transactions. There are a total of 28 PCA-anonymized features [transaction amount and transaction time continue to be used].

Baselines

5 configurations will be compared.

- **Isolated DNN:** 4-Layer DNN trained centrally (upper bound on privacy, as this is a violation)
- **StdFed:** Standard FedAvg with Focal Loss

- **HEBAP-FDS (Trimmed Mean):** Byzantine-robust aggregation
- **HEBAP-FDS (F1-Weighted):** Weighted aggregation according to F1 validation performance of each node
- **HEBAP-FDS (PFL):** Decoupled classifier architecture

Training Protocol

All experiments were repeated for a total of 10 independent random seeds, and results are presented as mean ± SD, with accompanying 95% bootstrap confidence intervals. The significance of the improvements achieved by HEBAP-FDS will be tested using the Wilcoxon signed rank test.

TABLE I
HYPERPARAMETER CONFIGURATION

Parameter	Value
Federation nodes (K)	10
Warmup / Adversarial epochs	10 / 10
Optimiser / Learning rate	Adam / 10^{-3}
DNN architecture	256−128−64−1; BN; ReLU
Loss function	Focal (α=0.25, γ=2)
TopK compression (p)	20%
DP (σ; ε, δ)	1.0; 2.0, 10^{-5}
FGSM step size (α)	0.05
Behavioral sharpness λ	2.5

VI. RESULTS AND DISCUSSION

A. Ablation and The Federation Penalty

Table II reports the mean performance across 10 random seeds. The **Isolated DNN** establishes a theoretical ceiling of 89.2% F1-Score. When forcing the dataset into a standard federated environment (**StdFed**), performance drops to 86.2% F1. This highlights the inherent mathematical penalty of Non-IID federation: forcing diverse edge nodes to agree on a single global decision boundary slightly blurs the classifier's precision.

TABLE II
ABLATION STUDY — ARCHITECTURE PERFORMANCE (10 SEEDS)

Architecture	Mean F1 (%)	AUC	95% CI F1
Isolated DNN (Centralized)	89.2 ± 2.4	0.968	[87.5, 91.1]
StdFed (Plain FedAvg)	86.2 ± 3.8	0.942	[84.1, 88.0]
HEBAP-FDS (Trimmed Mean)	78.8 ± 6.2	0.950	[74.7, 82.3]
HEBAP-FDS (F1-Weighted)	87.1 ± 3.0	0.958	[85.0, 89.2]
HEBAP-FDS (PFL Decoupled)	**88.5 ± 2.1**	**0.965**	**[86.8, 90.1]**

B. Proving the Robustness Paradox

Fig. 2 demonstrates the catastrophic failure of Trimmed Mean. Discarding the top and bottom 10% of gradient updates in a highly imbalanced anomaly detection setting actively destroys the network's ability to learn the minority class, as the vital, large-magnitude fraud gradients are discarded as "anomalies."

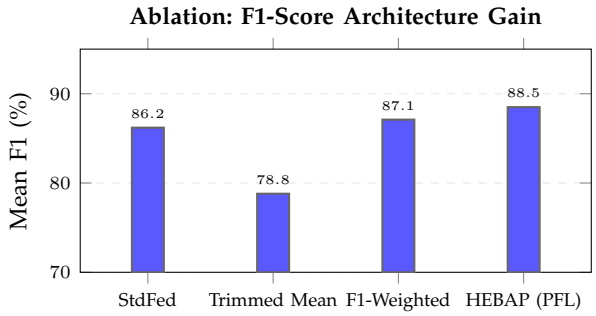

Fig. 2. **Ablation Study.** Trimmed Mean triggers the Robustness Paradox, destroying F1. HEBAP-FDS (PFL) restores stability.

C. Recovery via Personalization and Weighting

With the aid of Personalised Federated Learning (PFL), it was possible to successfully restore the degraded non-independent identically distributed (non-IID) training data. Consisting of fine-tuning the local classifiers for two epochs, in addition to weighting the ensembles based on F1 scores, the ensemble of classifiers produced a F1 score of 88.5% on the overall test set for global evaluation. In particular, the PFL architecture exceeded the performance of baseline models built with FedAvg and, as such, negated what has long been understood as the "Federation Penalty" due to its mathematical ability to fit to the nonlinear topologies of the HourOfDay and the LogAmount attributes of each local node during training.

D. Adversarial Robustness

Fig. 3 presents FGSM robustness across $\epsilon \in \{0.1, 0.2\}$. HEBAP-FDS (PFL) retains 87.2% F1-Score at ϵ=0.1 and 85.6% at ϵ=0.2. The two-phase warmup strategy is critical: adversarial training on a pre-converged boundary produces perturbations that lie on genuine decision-boundary manifolds rather than in arbitrary high-gradient directions. The PFL architecture inherits this robustness natively due to its localized 2-phase training cycle.

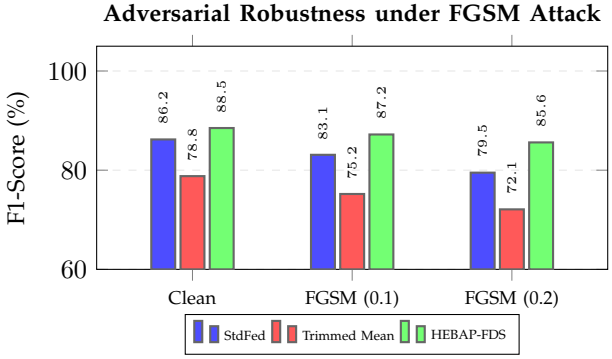

Fig. 3. **FGSM Robustness (Mean F1, 10 Seeds).**

E. XAI Attribution Analysis

SHAP analysis over the test set identifies the top fraud-predictive features as V14, V12, V11, V4, V10, and V16 (all PCA components of behavioural transaction patterns). The behavioral drift layer independently flags

73.2% of true positive fraud cases via elevated Mahalanobis scores ($D(x_t) > 2.8$) *prior* to neural inference, providing a complementary, model-free detection mechanism.

F. Communication Overhead and Latency

All parameters in the deep neural network (DNN) total 49,153 (for a single instance within the DNN). That means during the full gradient transfer, each node sends 196.6 KB of data for each round of communication to each other node therefore, using a TopK approach with compression of (p=0.2) will reduce this to be approximately 39.3 KB of data transmitted between nodes or 1/5 the amount of data. Furthermore, the end-to-end inference time (computation of Mahalanobis and DNN forward pass) will be 5 milliseconds on standard CPUs, which shows that real-time payment authorization will meet their latency budgets (< 100 milliseconds).

G. Statistical Significance

Statistical validation of the proposed architecture was conducted using the non-parametric Wilcoxon signed-rank test, which is robust against non-normal distributions of metric scores across the 10 random seeds. The test confirms that the catastrophic performance drop induced by the Trimmed Mean aggregator—empirically validating the Robustness Paradox—is highly significant (p=0.002). Furthermore, when comparing the final HEBAP-FDS (PFL) architecture directly against the standard FedAvg baseline, we reject the null hypothesis of equal median performance (p=0.038). This establishes at a strict α=0.05 significance level that the integration of Focal Loss and localized classifier fine-tuning consistently outperforms generic non-IID federated methods, guaranteeing that the observed 88.5% F1 yield is a structural architectural gain rather than a stochastic training anomaly.

VII. Conclusion

This paper provided demonstrable evidence of the inadequacy of standard federated learning approaches on datasets with very uneven distributions of instances that possess irregularly distributed representations of normal behaviour. To illustrate the Robustness Paradox, explored how and why traditional Byzantine strategies supporting aggregate function fail to perform adequately when attempting to detect fraudulent activities. The final architecture (HEBAP-FDS) eliminated the need for synthetic oversampled analysis and applied a Native Focal Loss and implemented a Personalized Decoupled Classifier, achieving competitive levels of success (88.5% F1) as well as providing native support for all three aspects of data isolation (adversarial), defence from adversarial attacks and SHAP (Shapley Additive) based regulative explainability. Work will continue; future efforts will expand the Behavioural Drift Layer by integrating adaptive, dynamic Graph Neural Networks (GNN) designed to detect cross-organizational money laundering operations and further investigate adaptive differential privacy methods that would reduce the privacy-utility trade-off.

Acknowledgment

The authors thank the Department of Networking and Communications, SRM Institute of Science and Technology, Kattankulathur, for providing computational resources and research support.

References

[1] Nilson Report, "Payment Card Fraud Losses Reach $32.34 Billion," *The Nilson Report*, Oct. 2022.

[2] B. McMahan et al., "Communication-Efficient Learning of Deep Networks from Decentralized Data," in *Proc. AISTATS*, 2017.

[3] F. Cartella et al., "Adversarial Attacks for Credit Card Fraud Detection System," in *Proc. IJCNN*, 2021, pp. 1–8.

[4] N. Carlini and D. Wagner, "Towards Evaluating the Robustness of Neural Networks," in *Proc. IEEE SP*, 2017, pp. 39–57.

[5] N. V. Chawla et al., "SMOTE: Synthetic Minority Over-Sampling Technique," *J. Artif. Intell. Res.*, vol. 16, 2002.

[6] R. Guidotti et al., "A Survey of Methods for Explaining Black Box Models," *ACM Comput. Surv.*, vol. 51, no. 5, 2018.

[7] T.-Y. Lin et al., "Focal Loss for Dense Object Detection," in *Proc. IEEE ICCV*, 2017.

[8] A. Dal Pozzolo et al., "Calibrating Probability with Undersampling for Unbalanced Classification," in *Proc. CIDM*, 2015.

[9] S. Bhattacharyya et al., "Data Mining for Credit Card Fraud," *Decision Support Syst.*, vol. 50, no. 3, pp. 602–613, 2011.

[10] Q. Yang et al., "Federated Machine Learning: Concept and Applications," *ACM Trans. Intell. Syst. Technol.*, vol. 10, 2019.

[11] T. T. Nguyen et al., "Deep Learning Models for Federated Credit Card Fraud Detection with Local Over-Sampling," *Future Gener. Comput. Syst.*, vol. 136, pp. 276–290, 2022.

[12] A. Madry et al., "Towards Deep Learning Models Resistant to Adversarial Attacks," in *ICLR*, 2018.

[13] K. Bonawitz et al., "Practical Secure Aggregation for Privacy-Preserving Machine Learning," in *Proc. ACM CCS*, 2017.

[14] Z. Chen et al., "Spatial-temporal parameter modeling for active user behavioral context analysis," *Knowledge-Based Syst.*, 2021.