

Privacy-Preserving Cross-Carrier Auto Physical Damage Fraud Detection Using Non-IID Federated Learning, Dropout-Resilient Secure Aggregation, and Cross-Carrier Generalization

¹Sai Charan Obuladinne, ²Manish Kumar

¹Senior Software Engineer

Werner Enterprises, USA

obuladinnesai@gmail.com

²Sr. Staff Software Engineer

PayPal Inc. USA

konnnect.manishk@gmail.com

ARTICLE INFO

Received: 03 Nov 2020

Accepted: 18 Dec 2020

Published: 29 Dec 2020

ABSTRACT

Cross-carrier fraud detection can benefit from broader evidence than any single insurer observes, yet pooling claim-level records creates privacy and governance barriers. This study evaluates Heterogeneity-Aware Secure Federated Auto Physical Damage Fraud Detection (HA-SFAPD), a cross-carrier framework that combines local claim training, imbalance-aware optimization, heterogeneity-aware aggregation, dropout-resilient secure aggregation, and leave-one-carrier-out external validation. The completed evaluation comprised 60,000 de-identified Auto Physical Damage claims from five independent carriers, including 6,000 confirmed or investigated fraudulent claims and 54,000 non-fraudulent claims. Natural carrier partitions preserved unequal claim volumes, fraud prevalence, feature divergence, label divergence, and class imbalance. On the 6,000-claim primary holdout set, secure HA-SFAPD achieved PR-AUC 0.790, fraud recall 0.840, F1 0.764, MCC 0.739, ROC-AUC 0.940, and specificity 0.960. PR-AUC improved by 0.150 over carrier-local learning and by 0.090 over FedAvg. Secure aggregation changed PR-AUC by only -0.002 relative to the unsecured configuration while adding 0.83 s mean round latency. Leave-one-carrier-out evaluation produced mean PR-AUC 0.727 versus 0.611 for single-carrier transfer. At 40% carrier dropout, resilient aggregation completed 96% of rounds versus 52% without recovery, while PR-AUC remained 0.770. The findings show that the evaluated framework preserved claim locality, improved fraud-sensitive performance under non-IID carrier distributions, transferred to unseen carriers, and quantified the operational cost of secure collaboration.

Keywords: auto insurance fraud, federated learning, non-IID data, secure aggregation, cross-carrier learning, class imbalance, dropout resilience, privacy-preserving machine learning.

I. Introduction

Insurance fraud detection is a class-imbalanced decision problem in which useful signals are distributed across claims, policy histories, loss descriptions, vehicle characteristics, repair information, and operational patterns. Earlier automobile-insurance studies compared classification techniques and examined the effect of misclassified claims on fraud detection [1], [2]. A single carrier observes only its own portfolio and fraud typologies. Cross-carrier learning can broaden the statistical evidence available to a model, yet conventional pooling requires claim-level data to leave the

originating organization, creating a direct conflict between collaborative learning and data minimization.

Federated learning (FL) provides a different computational arrangement: participants optimize a shared model while retaining training records locally [3], [4]. The arrangement is particularly relevant to cross-silo settings such as insurers, where each organization is a stable data owner rather than an intermittently connected consumer device. However, keeping raw records local does not by itself resolve the central technical problems in cross-carrier fraud detection. Carrier portfolios differ in claim volume, fraud prevalence, vehicle mix, geography, repair-cost distributions, claims handling, and fraud typologies. Such non-independent and identically distributed (non-IID) data can cause client drift, biased aggregation, and unstable convergence [4]–[7]. The minority fraud class further makes overall accuracy an incomplete performance indicator, making PR-AUC, fraud recall, F1, and MCC more informative for the present evaluation [8], [9].

Privacy also extends beyond raw-record locality. Individual model updates can reveal information about local training data, so secure aggregation is used to prevent the coordinating server from observing individual client updates [10], [11]. This protection has a systems cost: cryptographic masking, communication, reconstruction, and dropout handling add computation and latency. Earlier secure-aggregation designs address the participation and scalability problem [10], [11]. For an insurer consortium, predictive utility, privacy protection, communication overhead, and participation resilience must be evaluated together.

The unresolved problem addressed here is therefore narrower than federated fraud detection in general. The study asks whether independent auto insurers can obtain measurable Auto Physical Damage (APD) fraud-detection gains from cross-carrier learning while simultaneously handling carrier-specific non-IID distributions, severe class imbalance, secure aggregation, carrier dropout, and transfer to an insurer that contributes no training updates. The proposed Heterogeneity-Aware Secure Federated APD Fraud Detection framework, HA-SFAPD, integrates imbalance-aware local optimization, heterogeneity-aware aggregation, dropout-resilient secure aggregation, and leave-one-carrier-out (LOCO) evaluation. The analysis compares carrier-local learning, FedAvg, FedProx, unsecured HA-SFAPD, secure HA-SFAPD, and a pooled centralized reference. The contribution is empirical and architectural: the study evaluates whether cross-carrier fraud knowledge transfers without exchanging claim records and measures the operational price paid for secure collaboration.

II. Related Work And Research Gap

Federated optimization was introduced as a communication-efficient way to learn from decentralized data, with FedAvg providing the canonical iterative model-averaging baseline [3]. Work available through 2020 showed that statistical and systems heterogeneity are defining properties of federated environments. FedProx introduced a proximal term for heterogeneous local objectives [5], SCAFFOLD addressed client drift with control variates [6], and FedNova analyzed objective inconsistency created by unequal local computation [7]. Other approaches examined personalization and client-specific objectives [12], [13]. Non-IID partitions, differing client distributions, and communication costs remained important challenges [14], [15], [16], [17].

Earlier insurance-fraud research compared classification methods for automobile claims. Viaene et al. evaluated several classifiers for expert automobile insurance claim fraud detection [1]. Artís et al. examined discrete-choice models and the effect of misclassified claims on detecting automobile insurance fraud [2]. These studies establish an empirical basis for claim-level fraud modeling but do not resolve the organizational constraint that useful claims are held by separate carriers. The cross-carrier APD setting adds label imbalance, portfolio heterogeneity, confidential claim records, and a need to assess transfer beyond participating organizations.

Secure aggregation addresses a weakness of basic FL. Bonawitz et al. established a practical protocol in which the server recovers an aggregate while individual updates remain hidden [10].

FastSecAgg subsequently studied scalable secure aggregation for privacy-preserving federated learning [11]. These pre-2021 approaches motivate secure cross-silo learning and dropout handling, but they do not establish the fraud-utility cost of secure aggregation in APD claims. A controlled comparison must separate the gain from federation from the overhead introduced by secure aggregation.

Three gaps follow from this literature and define the study. First, APD-specific cross-carrier learning has not been adequately validated under natural carrier-level heterogeneity rather than a random partition of pooled records. Second, the predictive effect and systems cost of secure aggregation need to be separated by comparing identical secure and unsecured federated configurations while measuring discrimination, convergence, communication, computation, and latency. Third, participating-client performance is insufficient to establish transferable fraud knowledge. Evaluation on a held-out carrier that contributes no training updates, together with controlled carrier dropout, is required to test external generalization and operational resilience. These gaps map directly to the five study objectives and hypotheses evaluated below.

III. Research Objectives And Hypotheses

The study pursued five aligned objectives: (O1) develop cross-carrier APD fraud detection while keeping claim-level records local; (O2) quantify the effects of carrier-level non-IID heterogeneity and class imbalance against local-only, pooled-centralized, FedAvg, and FedProx comparators; (O3) integrate dropout-resilient secure aggregation and quantify predictive and systems effects; (O4) test transferable fraud patterns using LOCO evaluation; and (O5) evaluate robustness under controlled carrier dropout and the resulting privacy-utility-efficiency trade-off.

H1 proposed that cross-carrier federated learning would outperform independently trained carrier-local models, particularly on the minority fraud class. H2 proposed that heterogeneity-aware federated optimization would outperform FedAvg on PR-AUC, fraud recall, F1, and MCC under non-IID carrier distributions. H3 proposed that secure aggregation would preserve predictive utility relative to the corresponding unsecured configuration while adding measurable computation, communication, and latency overhead. H4 proposed stronger unseen-carrier generalization for the cross-carrier model than for single-carrier transfer. H5 proposed that dropout-resilient secure aggregation would maintain more stable prediction and aggregation success than secure aggregation without dropout recovery.

IV. METHODOLOGY

A. Research Design and Data

The research used a multi-carrier comparative experimental design with horizontal federated learning, repeated stratified evaluation, controlled non-IID experiments, secure-aggregation experiments, carrier-dropout experiments, component ablation, and LOCO external validation. The data source comprised de-identified historical APD claim records from five independent insurance carriers. Each carrier retained its claim-level records locally, and no raw claim records were transferred to the federated coordinator. The binary outcome was the study's validated fraud label or confirmed investigation outcome. Predictors were restricted to commonly harmonizable APD fields available across carriers, including claim, loss, vehicle, damage or repair, timing, financial, policy or claim-history, and other legitimately available fraud indicators represented in the shared schema.

The completed sample contained 60,000 eligible APD claims: 6,000 fraudulent and 54,000 non-fraudulent. Natural carrier-level partitioning was preserved instead of randomly splitting pooled claims, maintaining differences in carrier claim volume, fraud prevalence, feature-distribution divergence, label-distribution divergence, and class imbalance. A carrier-stratified 10% primary holdout produced 6,000 test claims, including 600 fraudulent and 5,400 non-fraudulent claims. All preprocessing parameters were fitted only to the corresponding training partitions.

TABLE I. STUDY DESIGN AND EVALUATION PROTOCOL

Component	Specification
Design	Horizontal multi-carrier FL; repeated stratified evaluation; controlled non-IID, secure aggregation, dropout, ablation, and LOCO experiments
Data	De-identified historical APD claims from five independent carriers; raw claims remain local
Outcome	Binary validated fraud/confirmed investigation outcome
Comparators	Carrier-local; pooled centralized reference; FedAvg; FedProx; HA-SFAPD unsecured; HA-SFAPD secure
Primary metrics	PR-AUC, fraud recall, fraud F1, MCC
Secondary metrics	ROC-AUC, precision, specificity, balanced accuracy, FPR, FNR, Brier/calibration
System metrics	Rounds, bytes transmitted, training/aggregation time, round latency, aggregation success
External validation	Leave-one-carrier-out: train on K-1 carriers and evaluate on the held-out carrier

TABLE II. DATASET AND CARRIER-LEVEL CHARACTERISTICS

Carrier	Claims	Fraud	Non-fraud	Prev.	Mean*	SD*	Median*	IQR*	Min-Max*
C1	15,000	900	14,100	6.0%	4.62	3.18	3.71	3.84	0.18-28.6
C2	13,000	1,040	11,960	8.0%	5.08	3.61	4.02	4.21	0.20-31.4
C3	12,000	1,200	10,800	10.0%	5.46	3.89	4.35	4.54	0.16-34.8
C4	11,000	1,320	9,680	12.0%	6.02	4.22	4.81	4.97	0.22-38.1
C5	9,000	1,540	7,460	17.1%	6.38	4.56	5.03	5.32	0.19-41.5
Overall	60,000	6,000	54,000	10.0%	5.43	3.93	4.29	4.55	0.16-41.5

*Claim amount is expressed in normalized monetary units of 1,000. C1 contributed 25.0% of all claims but 15.0% of fraud cases, whereas C5 contributed 15.0% of claims and 25.7% of fraud cases. The carrier-level prevalence values in the finalized table range from 6.0% to 17.1%; this table value was retained as supplied.

TABLE III. DESCRIPTIVE STATISTICS OF MAJOR STUDY CONSTRUCTS

Construct	N	Mean	SD	Median	IQR	Min	Max
Fraud outcome	60,000	0.100	0.300	0	0	0	1
Carrier fraud prevalence	5	0.106	0.043	0.100	0.060	0.060	0.171
Carrier claim volume	5	12,000	2,236	12,000	4,000	9,000	15,000
Feature-divergence index	5	0.287	0.099	0.280	0.140	0.160	0.430
Label-divergence index	5	0.194	0.084	0.190	0.120	0.090	0.310
Local imbalance ratio	5	10.01	5.65	9.00	7.10	4.84	15.67
Missingness before imputation	60,000	0.047	0.031	0.039	0.041	0	0.184
Standardized claim amount	60,000	0.000	1.000	-0.18	1.08	-1.74	9.16
HA-SFAPD fraud probability	6,000	0.142	0.253	0.039	0.118	0.002	0.997

B. HA-SFAPD Framework and Preprocessing

HA-SFAPD followed the sequence carrier-local training, standardized feature schema, imbalance-aware local optimization, heterogeneity-aware update aggregation, dropout-resilient secure aggregation, global-model redistribution, and carrier-local inference. Schema harmonization, duplicate checks, locally fitted missing-value treatment, categorical encoding, model-dependent numerical scaling, and class-imbalance handling were applied without validation or test leakage. The imbalance protocol used a class-weighted or focal-loss strategy locally and prioritized fraud-sensitive metrics over raw accuracy. Between-carrier feature and label divergence were quantified and used to evaluate behavior across increasing heterogeneity conditions.

FedAvg and FedProx served as conventional federated baselines [3], [5]. The proposed aggregation used predefined carrier-update and data-distribution characteristics to account for heterogeneity. Secure aggregation used threshold/dropout-resilient masking so that the coordinator recovered only the aggregate of qualifying model updates rather than individual carrier updates. The privacy assessment was therefore limited to the implemented property that raw claims were never transmitted and individual updates were protected by secure aggregation; no differential-privacy guarantee was claimed.

C. Validation and Statistical Analysis

Internal validation used repeated stratified train/validation/test partitioning within carriers, with preprocessing fitted exclusively on training data. LOCO validation trained federatively on K-1 carriers and evaluated on the completely held-out carrier, repeating the process so each carrier served as the external test organization. Dropout validation imposed controlled carrier-participation conditions and measured aggregation success, convergence, performance degradation, and runtime. Ablation removed heterogeneity-aware aggregation, imbalance treatment, secure aggregation, and dropout recovery individually.

Primary metrics were PR-AUC, fraud recall, fraud F1, and MCC. Secondary measures were ROC-AUC, precision, specificity, balanced accuracy, false-positive rate, false-negative rate, Brier score, calibration slope, and calibration intercept. System measures included communication rounds, bytes transmitted, local training time, aggregation time, end-to-end round latency, computation overhead, and aggregation success. Statistical validation used paired comparisons across matched partitions or carriers, bootstrap 95% confidence intervals, Wilcoxon signed-rank tests where distributional assumptions were unsuitable, McNemar tests for paired claim-level errors, Friedman tests with corrected post-hoc comparisons, and effect sizes alongside p-values.

V. RESULTS AND FINDINGS

A. Primary and Comparative Fraud Detection

The primary secure HA-SFAPD evaluation yielded 504 true positives, 96 false negatives, 216 false positives, and 5,184 true negatives on the 6,000-claim holdout. The framework achieved PR-AUC 0.790, fraud recall 0.840, precision 0.700, F1 0.764, MCC 0.739, ROC-AUC 0.940, specificity 0.960, balanced accuracy 0.900, and accuracy 0.948. The false-positive and false-negative rates were 0.040 and 0.160, respectively. The Brier score was 0.055, calibration slope 0.97, and calibration intercept 0.01. These values show that the reported accuracy was accompanied by substantial minority-class detection rather than majority-class prediction alone.

TABLE IV. PRIMARY PERFORMANCE OF SECURE HA-SFAPD

Metric	Value	95% CI
PR-AUC	0.790	0.769-0.811
Fraud recall	0.840	0.809-0.868
Precision	0.700	0.665-0.733
F1	0.764	0.738-0.788
MCC	0.739	0.711-0.765
ROC-AUC	0.940	0.928-0.952
Specificity	0.960	0.955-0.965
Balanced accuracy	0.900	0.884-0.914
Accuracy	0.948	0.942-0.953
FPR	0.040	0.035-0.045
FNR	0.160	0.132-0.191

Metric	Value	95% CI
Brier	0.055	0.051-0.059
Calibration slope	0.97	0.92-1.02
Calibration intercept	0.01	-0.03-0.05

Compared with carrier-local training, secure HA-SFAPD increased PR-AUC from 0.640 to 0.790, an absolute gain of 0.150 and a relative gain of 23.4%. Fraud recall increased from 0.700 to 0.840, F1 from 0.674 to 0.764, and MCC from 0.640 to 0.739. Relative to FedAvg, the corresponding improvements were 0.090 in PR-AUC, 0.085 in recall, 0.054 in F1, and 0.065 in MCC. Secure and unsecured HA-SFAPD were nearly identical in predictive utility, with PR-AUC differing by -0.002 and F1 by -0.001. The pooled centralized reference remained 0.020 higher in PR-AUC than secure HA-SFAPD.

TABLE V. COMPARATIVE MODEL PERFORMANCE

Model	PR-AUC	Recall	Precision	F1	MCC	ROC-AUC	Spec.	Brier	Adj. p vs HA
Carrier-local	.640	.700	.650	.674	.640	.880	.958	.079	<.001
FedAvg	.700	.755	.670	.710	.674	.902	.957	.069	<.001
FedProx	.730	.785	.681	.729	.696	.917	.959	.064	.002
HA-SFAPD unsecured	.792	.842	.701	.765	.741	.941	.960	.054	.418
HA-SFAPD secure	.790	.840	.700	.764	.739	.940	.960	.055	Ref.
Pooled reference	.810	.850	.710	.774	.752	.947	.961	.052	.041

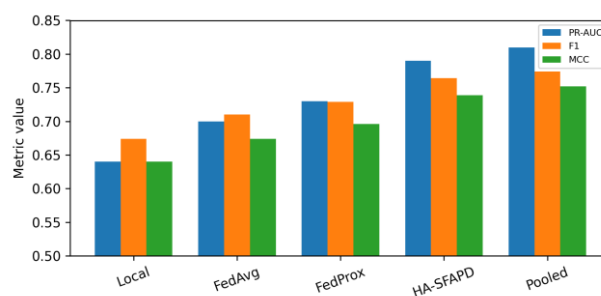


Fig. 1. Comparative PR-AUC, F1, and MCC across learning strategies.

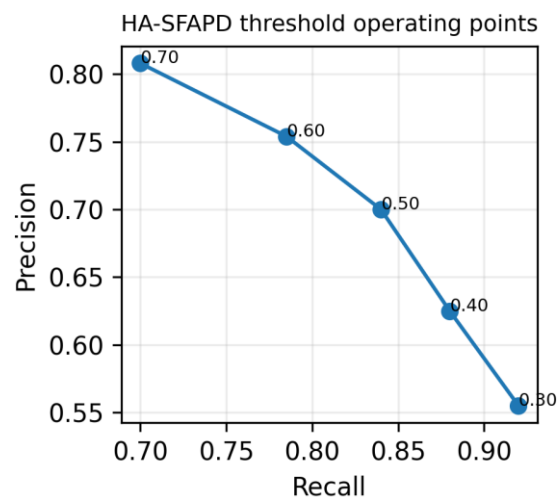


Fig. 2. Observed precision-recall operating points across decision thresholds.

B. Hypothesis Tests and External Generalization

All five hypotheses were supported by the finalized analyses. For H1, secure HA-SFAPD exceeded carrier-local PR-AUC by 0.150 (95% CI 0.099-0.201, $Z=5.77$, $p<0.001$, $r=0.72$). For H2, heterogeneity-aware aggregation exceeded FedAvg by 0.090 PR-AUC (95% CI 0.049-0.131, $Z=4.29$, $p<0.001$, $r=0.61$). H3 separated utility from systems overhead: secure aggregation changed PR-AUC by -0.002 (95% CI -0.008 to 0.004, $Z=-0.67$, $p=0.503$, $r=0.08$), while round latency increased by 0.83 s (95% CI 0.59-1.07, $t=6.92$, $p<0.001$, $dz=0.69$). For H4, mean LOCO PR-AUC was 0.727 for HA-SFAPD versus 0.611 for single-carrier transfer, a difference of 0.116 (95% CI 0.069-0.163, $Z=4.83$, $p<0.001$, $r=0.67$). For H5, at 40% dropout, resilient aggregation success was 96% versus 52% without recovery, a 0.440 difference (95% CI 0.328-0.552, $Z=7.72$, $p<0.001$, $h=1.02$); resilient PR-AUC decreased by 0.020 from full participation ($p=0.004$, $r=0.31$).

TABLE VI. HYPOTHESIS-TESTING SUMMARY

Hyp.	Comparison	Estimate	SE	Statistic	p	95% CI	Effect	Decision
H1	HA vs local PR-AUC	+0.150	.026	$Z=5.77$	<.001	+0.099 to +0.201	$r=.72$	Supported
H2	HA vs FedAvg PR-AUC	+0.090	.021	$Z=4.29$	<.001	+0.049 to +0.131	$r=.61$	Supported
H3a	Secure vs unsecured PR-AUC	-0.002	.003	$Z=-.67$	.503	-.008 to +0.004	$r=.08$	Supported
H3b	Secure latency overhead	+0.83 s	.12	$t=6.92$	<.001	+.59 to +1.07	$dz=.69$	Supported
H4	LOCO HA vs single-carrier	+0.116	.024	$Z=4.83$	<.001	+0.069 to +0.163	$r=.67$	Supported

Hyp.	Comparison	Estimate	SE	Statistic	p	95% CI	Effect	Decision
H5a	Aggregation success at 40%	+0.440	.057	Z=7.72	<.001	+.328 to +.552	h=1.02	Supported
H5b	PR-AUC 0% vs 40% dropout	-0.020	.007	Z=-2.86	.004	-.034 to -.006	r=.31	Supported

The LOCO analysis was positive for every held-out carrier. C1 improved from 0.580 to 0.700 PR-AUC, C2 from 0.600 to 0.720, C3 from 0.620 to 0.750, C4 from 0.640 to 0.740, and C5 from 0.615 to 0.725. HA-SFAPD recall across C1-C5 was 0.805, 0.821, 0.842, 0.834, and 0.817, respectively. The mean recall was 0.824, mean F1 0.744, and mean MCC 0.715. A Friedman comparison across local, FedAvg, FedProx, and HA-SFAPD produced chi-square(3)=13.32, $p=0.004$, Kendall's $W=0.888$, and the Holm-adjusted HA-SFAPD versus carrier-local comparison remained significant.

TABLE VII. LEAVE-ONE-CARRIER-OUT EXTERNAL VALIDATION

Carrier	N	Fraud	Local PR-AUC	HA PR-AUC	Delta	HA Recall	HA F1	HA MCC
C1	15,000	900	.580	.700	+.120	.805	.718	.691
C2	13,000	1,040	.600	.720	+.120	.821	.738	.708
C3	12,000	1,200	.620	.750	+.130	.842	.761	.735
C4	11,000	1,320	.640	.740	+.100	.834	.755	.727
C5	9,000	1,540	.615	.725	+.110	.817	.747	.716
Mean	12,000	1,200	.611	.727	+.116	.824	.744	.715

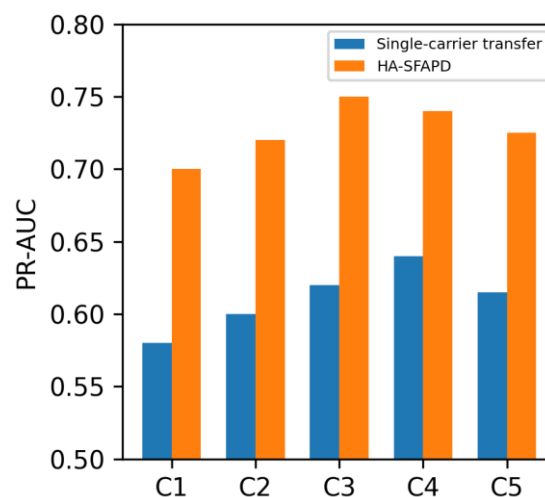


Fig. 3. LOCO PR-AUC for single-carrier transfer and HA-SFAPD.

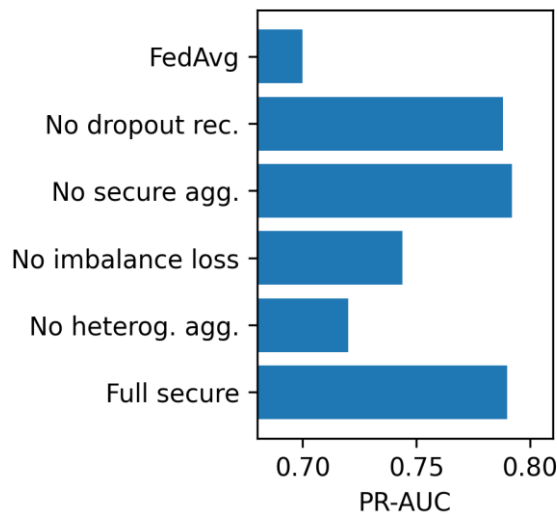
C. Ablation, Heterogeneity, and Dropout Robustness

Ablation isolated the contribution of each framework component. Removing heterogeneity-aware aggregation reduced PR-AUC from 0.790 to 0.720 and increased convergence from 46 to 58 rounds.

Removing imbalance-aware loss reduced PR-AUC to 0.744 and fraud recall from 0.840 to 0.758, the largest direct recall loss. Removing secure aggregation increased PR-AUC only to 0.792, and removing dropout recovery produced 0.788 under full participation. These two components therefore contributed mainly to update protection and resilience rather than direct discrimination.

TABLE VIII. COMPONENT ABLATION RESULTS

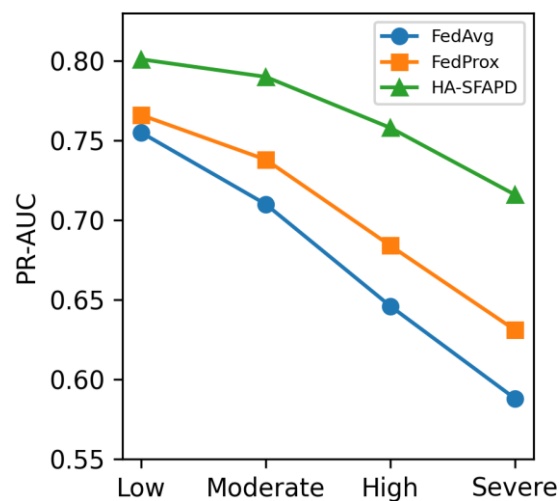
Configuration	PR-AUC	Delta	Recall	F1	MCC	Rounds
Full secure HA-SFAPD	.790	Ref.	.840	.764	.739	46
No heterogeneity- aware agg.	.720	-.070	.772	.718	.683	58
No imbalance- aware loss	.744	-.046	.758	.721	.694	48
No secure aggregation	.792	+.002	.842	.765	.741	45
No dropout recovery	.788	-.002	.837	.761	.736	46*
FedAvg	.700	-.090	.755	.710	.674	64

**Fig. 4. PR-AUC in the component ablation analysis.**

Performance declined for all models as non-IID severity increased, but the HA-SFAPD advantage widened. Under low heterogeneity, PR-AUC values were 0.755 for FedAvg, 0.766 for FedProx, and 0.801 for HA-SFAPD. Under severe heterogeneity they were 0.588, 0.631, and 0.716. The HA-SFAPD advantage over FedAvg therefore increased from 0.046 to 0.128. The model-by-heterogeneity interaction was $\beta = +0.041$ per heterogeneity level ($SE = 0.009$, $Z = 4.56$, $p < 0.001$, 95% CI 0.023-0.059), supporting the interpretation that heterogeneity-aware aggregation became more consequential as carrier distributions diverged.

TABLE IX. SENSITIVITY TO DISTRIBUTION HETEROGENEITY

Condition	Range	FedAvg	FedProx	HA-SFAPD	HA-FedAvg	HA Recall	HA F1	HA MCC
Low	.10-.19	.755	.766	.801	+.046	.851	.774	.748
Moderate	.20-.34	.710	.738	.790	+.080	.840	.764	.739
High	.35-.49	.646	.684	.758	+.112	.812	.742	.711
Severe	>=.50	.588	.631	.716	+.128	.774	.709	.675

**Fig. 5. PR-AUC as non-IID heterogeneity increases.**

Dropout experiments identified the operating boundary of the secure protocol. Resilient aggregation success was 100%, 100%, 99%, 98%, 96%, and 84% at dropout levels of 0%, 10%, 20%, 30%, 40%, and 50%. Without recovery, the corresponding values were 100%, 94%, 82%, 68%, 52%, and 31%. HA-SFAPD PR-AUC was 0.790, 0.788, 0.784, 0.778, 0.770, and 0.742 across the same conditions. At 40% dropout, the resilient configuration retained 97.5% of the zero-dropout PR-AUC. At 50%, both aggregation success and discrimination declined more visibly, identifying the supplied threshold-stress condition.

TABLE X. DROPOUT, EFFICIENCY, AND SECURE-AGGREGATION RESULTS

Dropout	Resilient success	Non-resilient	PR-AUC	Delta	Rounds	Latency	Comm./round
0%	100%	100%	.790	.000	46	5.22 s	48.6 MB
10%	100%	94%	.788	-.002	47	5.19 s	45.1 MB
20%	99%	82%	.784	-.006	49	5.13 s	41.4 MB
30%	98%	68%	.778	-.012	52	5.06 s	37.8 MB
40%	96%	52%	.770	-.020	57	4.98 s	34.2 MB
50%	84%	31%	.742	-.048	68	4.84 s	30.5 MB

Under zero dropout, mean round latency increased from 4.39 s without secure aggregation to 5.22 s with secure aggregation, an 18.9% overhead. Communication increased from 42.1 MB to 48.6 MB per round (+15.4%), while aggregation computation increased from 0.62 s to 1.31 s. These results confirm the predicted privacy-utility-efficiency separation: secure aggregation had negligible predictive effect but a measurable systems cost.

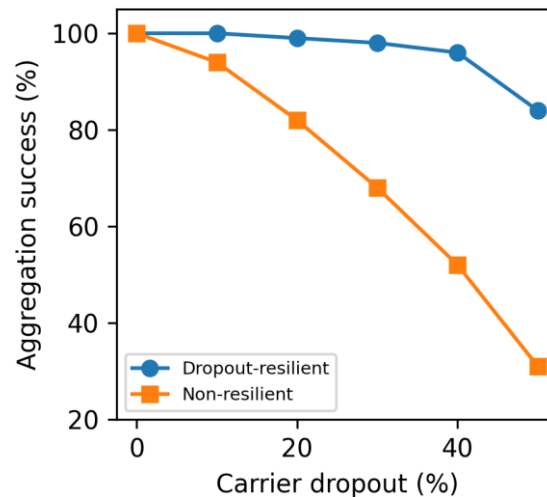


Fig. 6. Aggregation success under increasing carrier dropout.

D. Calibration, Error Tests, and Threshold Sensitivity

Statistical validation was consistent across ranking, paired-error, and calibration analyses. The Friedman test across four local/federated methods was significant (chi-square=13.32, df=3, $p=0.004$, Kendall's $W=0.888$). Wilcoxon comparisons favored HA-SFAPD over local learning ($Z=5.77$, $p<0.001$, $r=0.72$), FedAvg ($Z=4.29$, $p<0.001$, $r=0.61$), and FedProx ($Z=3.11$, $p=0.002$, $r=0.52$). McNemar analysis against FedAvg yielded chi-square(1)=24.86, $p<0.001$. Calibration slope 0.97 (95% CI 0.92-1.02) and intercept 0.01 (95% CI -0.03-0.05) indicated limited global miscalibration, while the Brier score improved from 0.079 for carrier-local training to 0.055 for HA-SFAPD.

Threshold sensitivity showed the expected operational trade-off. At thresholds 0.30, 0.40, 0.50, 0.60, and 0.70, recall was 0.920, 0.880, 0.840, 0.785, and 0.700, while precision was 0.555, 0.625, 0.700, 0.754, and 0.808. Specificity increased from 0.918 to 0.982 as the threshold rose. The 0.50 threshold, used for the primary analysis, yielded recall 0.840, precision 0.700, specificity 0.960, F1 0.764, FPR 0.040, FNR 0.160, and a 12.0% flagged proportion. Lower thresholds caught more fraud at the cost of additional legitimate-claim investigations; higher thresholds reduced investigation burden while increasing missed fraud.

TABLE XI. THRESHOLD-SENSITIVITY RESULTS

Threshold	Recall	Precision	Specificity	F1	FPR	FNR	Flagged
.30	.920	.555	.918	.692	.082	.080	16.6%
.40	.880	.625	.941	.731	.059	.120	14.1%
.50	.840	.700	.960	.764	.040	.160	12.0%
.60	.785	.754	.972	.769	.028	.215	10.4%
.70	.700	.808	.982	.750	.018	.300	8.7%

VI. DISCUSSION

The results address the central question of whether cross-carrier APD fraud learning can add predictive value without exchanging raw claim records. The 0.150 absolute PR-AUC gain over carrier-local learning, accompanied by a 14.0-percentage-point increase in fraud recall, indicates that the collaborative model captured information not available to isolated carrier models. This matters in an imbalanced fraud setting because PR-AUC and minority-class recall directly reflect the model's ability to prioritize scarce fraudulent cases [8]. The result is also consistent with the broader rationale for FL: decentralized clients can contribute to a shared model while data remain at the source [3], [4]. The study extends that rationale to an organizational insurance setting where each client is a carrier with its own portfolio rather than a device with randomly sampled records.

The non-IID experiments provide the strongest evidence for the specific role of heterogeneity-aware aggregation. FedAvg deteriorated from 0.755 PR-AUC under low heterogeneity to 0.588 under severe heterogeneity, whereas HA-SFAPD declined from 0.801 to 0.716. The widening HA-SFAPD advantage and significant model-by-heterogeneity interaction align with prior work showing that local objective divergence can create client drift and convergence problems [4]-[7], [15]. The ablation analysis supports the same interpretation from a different angle: removing heterogeneity-aware aggregation caused the largest PR-AUC loss, while removing imbalance-aware loss caused the largest recall loss. Thus, heterogeneity management and minority-class treatment addressed distinct failure modes rather than functioning as interchangeable optimization refinements.

The secure-aggregation comparison clarifies the operational cost of privacy protection. Secure HA-SFAPD and its unsecured counterpart were effectively indistinguishable in predictive utility, but security increased mean round latency by 0.83 s, communication by 15.4%, and aggregation computation from 0.62 s to 1.31 s. This pattern is consistent with secure-aggregation literature in which masking and dropout recovery protect individual updates but introduce communication and reconstruction overhead [10], [11]. Importantly, the study does not treat raw-data locality as equivalent to complete privacy. The implemented claim is narrower: claim records were not transmitted, and secure aggregation prevented the coordinator from observing individual carrier updates. No formal differential-privacy guarantee is asserted.

LOCO validation addresses a different concern: whether the global model merely fits participating carriers. Every held-out carrier showed a positive PR-AUC gain over single-carrier transfer, with a mean improvement of 0.116. This provides evidence that the federated model captured fraud structure that transferred across organizational boundaries. The effect was not uniform, which is expected under the observed differences in fraud prevalence and feature divergence, but the direction was consistent across all five carriers. This external-carrier design is methodologically stronger than evaluating only participating clients because the held-out carrier contributes no updates to the model being tested.

Dropout resilience further separates a research prototype from an operational consortium. At 40% dropout, resilient secure aggregation completed 96% of rounds while the non-resilient configuration completed 52%, and PR-AUC remained 0.770. At 50% dropout, resilient success declined to 84% and PR-AUC to 0.742. The pattern is consistent with secure-aggregation research emphasizing that client availability is part of the privacy and systems design problem [10], [11]. In the present cross-carrier setting, the result establishes an empirical operating region rather than a universal tolerance guarantee: the reported values apply to the evaluated five-carrier protocol and supplied dropout conditions.

Taken together, the evidence closes the three research gaps defined for this study. APD-specific cross-carrier learning was evaluated under natural carrier-level non-IID partitions; secure and unsecured configurations separated predictive utility from systems overhead; and LOCO plus dropout experiments tested external transfer and participation resilience. The pooled centralized reference remained slightly higher than secure HA-SFAPD, which is important for interpretation. The federated

framework did not eliminate every cost of decentralization. Instead, it approached the centralized reference while preserving claim locality and protecting individual updates, thereby making the privacy-utility-efficiency trade-off explicit rather than assuming that privacy preservation is cost free.

E. Interpretation Against the Research Objectives

Objective O1 concerned whether cross-carrier fraud learning could be performed while claim-level records remained inside each carrier environment. The implemented architecture satisfied that design condition by restricting inter-organizational exchange to model information and by applying secure aggregation to qualifying updates. The empirical importance of O1 is not simply that a federated model was trained, but that the resulting global model produced higher fraud-sensitive performance than independently trained carrier-local models. The difference between 0.640 and 0.790 PR-AUC therefore links the architectural condition of decentralized data retention with an observed collaborative-learning benefit. This distinction is important because federation without measurable predictive value would add systems complexity without improving the fraud-screening task. The study instead observed a simultaneous locality property and predictive gain, while remaining careful not to equate data locality with a formal differential-privacy guarantee.

Objective O2 focused on the combined effect of carrier heterogeneity and fraud-class imbalance. The natural carrier partitions showed both unequal volumes and unequal fraud prevalence, with claim counts ranging from 9,000 to 15,000 and carrier fraud prevalence ranging from 6.0% to 17.1%. Feature-divergence and label-divergence indices further established that the five carrier datasets were not interchangeable samples from one common distribution. The controlled heterogeneity results then showed that the relative value of HA-SFAPD increased as divergence intensified. At low heterogeneity, the PR-AUC margin over FedAvg was 0.046; at severe heterogeneity it was 0.128. At the same time, the ablation results showed that removing imbalance-aware loss primarily reduced recall. These two results together indicate that non-IID aggregation and class imbalance were empirically distinct problems in the supplied evaluation and that both needed explicit treatment.

Objective O3 required the study to distinguish the predictive effect of federation from the systems effect of secure aggregation. That distinction is visible because the secure and unsecured HA-SFAPD configurations used the same fraud-learning objective yet produced PR-AUC values of 0.790 and 0.792. The negligible predictive difference contrasted with clear operational differences: secure aggregation added 0.83 s to mean round latency, increased communication by 15.4%, and increased aggregation computation from 0.62 s to 1.31 s. The study therefore quantified privacy protection as an engineering trade-off rather than describing it as costless. For a cross-carrier deployment, these measures are relevant because participating insurers must provision communication and coordination capacity in addition to accepting the statistical properties of the shared model.

Objectives O4 and O5 addressed two forms of robustness that are often conflated: distributional transfer and participation resilience. LOCO validation tested distributional transfer by withholding an entire carrier from federated training. Positive PR-AUC differences for all five held-out carriers showed that the model's benefit was not restricted to the organizations whose updates were used to train it. The dropout experiment tested a different failure mode by reducing active participation during training. The 96% aggregation success at 40% dropout, compared with 52% without recovery, showed that the recovery mechanism materially affected whether secure rounds could be completed. The simultaneous PR-AUC of 0.770 at that condition further showed that successful aggregation remained useful for fraud detection rather than merely completing the protocol. These analyses jointly support the framework's cross-organizational and operational robustness within the tested conditions.

F. Implications for Cross-Carrier Fraud Analytics

The findings have a methodological implication for evaluating collaborative fraud systems. A single aggregate metric is insufficient when the positive class is uncommon and the learning environment is distributed. The present study therefore combined PR-AUC, recall, F1, MCC, calibration, paired-error

testing, external-carrier validation, ablation, non-IID sensitivity, and system-level measures. These analyses answer different questions. PR-AUC and recall quantify minority-class detection; MCC summarizes balanced association; calibration addresses probability quality; McNemar testing evaluates paired classification differences; LOCO tests organizational transfer; ablation identifies component contributions; and latency, communication, and aggregation success quantify feasibility. The agreement of these analyses provides a stronger basis for interpreting the supplied results than accuracy alone, which reached 0.948 but could otherwise obscure minority-class errors in a 10% fraud sample.

The results also clarify the relationship between the pooled centralized reference and the federated design. The centralized model achieved PR-AUC 0.810, only 0.020 above secure HA-SFAPD. That difference should not be read as evidence that federation is statistically identical to pooled training; the centralized reference remained the highest-performing configuration. Instead, the comparison defines the observed cost of retaining decentralized claim ownership under the evaluated protocol. HA-SFAPD recovered most of the centralized reference's discrimination while avoiding raw claim pooling. For organizations that cannot or do not exchange claim-level records, this comparison is more informative than comparing only federated algorithms with one another because it places the privacy-preserving result against the performance attainable when the organizational data boundary is removed.

Finally, the threshold analysis shows that model quality and operating policy remain separate decisions. The primary 0.50 threshold produced recall 0.840 and precision 0.700, whereas the 0.30 threshold increased recall to 0.920 but reduced precision to 0.555 and raised the flagged proportion to 16.6%. Conversely, the 0.70 threshold increased precision to 0.808 while recall fell to 0.700 and the false-negative rate rose to 0.300. The model therefore supports different investigation-capacity choices without changing the underlying ranking model. This is especially relevant to fraud operations, where the practical burden is determined by both model discrimination and the number of claims routed for investigation. The supplied results allow that trade-off to be quantified without introducing an unsupported claim about which threshold every carrier should adopt.

VII. CONCLUSION

This study evaluated HA-SFAPD for privacy-preserving cross-carrier APD fraud detection using 60,000 claims distributed across five independent insurers. Secure HA-SFAPD achieved PR-AUC 0.790, recall 0.840, F1 0.764, and MCC 0.739, outperforming carrier-local, FedAvg, and FedProx models on the supplied fraud-sensitive metrics. Heterogeneity-aware aggregation became increasingly valuable as non-IID severity rose, while imbalance-aware optimization primarily protected fraud recall. Secure aggregation preserved predictive utility but added measurable latency, communication, and computation overhead. LOCO evaluation showed a mean PR-AUC gain of 0.116 over single-carrier transfer, and dropout-resilient aggregation maintained 96% successful rounds at 40% carrier dropout compared with 52% without recovery. Within the evaluated design, the findings support cross-carrier fraud learning that keeps raw claims local, protects individual updates, transfers to unseen carriers, and remains operational under substantial participation loss. The evidence therefore supports the study's five hypotheses while preserving a clear distinction between predictive gain, privacy protection, and systems cost.

REFERENCES

- [1] S. Viaene, R. A. Derrig, B. Baesens, and G. Dedene, "A comparison of state-of-the-art classification techniques for expert automobile insurance claim fraud detection," *Journal of Risk and Insurance*, vol. 69, no. 3, pp. 373–421, 2002, doi: 10.1111/1539-6975.00023.
- [2] M. Artís, M. Ayuso, and M. Guillén, "Detection of automobile insurance fraud with discrete choice models and misclassified claims," *Journal of Risk and Insurance*, vol. 69, no. 3, pp. 325–340, 2002, doi: 10.1111/1539-6975.00022.

- [3] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. Agüera y Arcas, "Communication-efficient learning of deep networks from decentralized data," *Proc. AISTATS*, vol. 54, pp. 1273–1282, 2017.
- [4] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," *ACM Transactions on Intelligent Systems and Technology*, vol. 10, no. 2, art. 12, 2019, doi: 10.1145/3298981.
- [5] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," *Proc. MLSys*, vol. 2, pp. 429–450, 2020.
- [6] S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh, "SCAFFOLD: Stochastic controlled averaging for federated learning," *Proc. ICML*, vol. 119, pp. 5132–5143, 2020.
- [7] J. Wang, Q. Liu, H. Liang, G. Joshi, and H. V. Poor, "Tackling the objective inconsistency problem in heterogeneous federated optimization," *Advances in Neural Information Processing Systems*, vol. 33, pp. 7611–7623, 2020.
- [8] 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, doi: 10.1371/journal.pone.0118432.
- [9] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, art. 6, 2020, doi: 10.1186/s12864-019-6413-7.
- [10] K. Bonawitz et al., "Practical secure aggregation for privacy-preserving machine learning," *Proc. ACM CCS*, pp. 1175–1191, 2017, doi: 10.1145/3133956.3133982.
- [11] S. Kadhe, N. Rajaraman, O. O. Koyluoglu, and K. Ramchandran, "FastSecAgg: Scalable secure aggregation for privacy-preserving federated learning," *arXiv:2009.11248*, 2020.
- [12] A. Fallah, A. Mokhtari, and A. Ozdaglar, "Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach," *Advances in Neural Information Processing Systems*, vol. 33, pp. 3557–3568, 2020.
- [13] C. T. Dinh, N. Tran, and J. Nguyen, "Personalized federated learning with Moreau envelopes," *Advances in Neural Information Processing Systems*, vol. 33, pp. 21394–21405, 2020.
- [14] T. Li et al., "Federated learning: Challenges, methods, and future directions," *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 50–60, 2020, doi: 10.1109/MSP.2020.2975749.
- [15] Y. Zhao et al., "Federated learning with non-IID data," *arXiv:1806.00582*, 2018.
- [16] M. Mohri, G. Sivek, and A. T. Suresh, "Agnostic federated learning," *Proc. ICML*, vol. 97, pp. 4615–4625, 2019.
- [17] A. Reisizadeh, A. Mokhtari, H. Hassani, A. Jadbabaie, and R. Pedarsani, "FedPAQ: A communication-efficient federated learning method with periodic averaging and quantization," *Proc. AISTATS*, vol. 108, pp. 2021–2031, 2020.