Indonesian Journal of Electrical Engineering and Computer Science
Vol. 44, No. 1, October 2026, pp. 140~149
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v44.i1.pp140-149

Multi-dataset machine learning pipeline for insurance fraud detection

Somkiat Kosolsombat[1], Chiabwoot Ratanavilisagul[2]
[1]Data Science and Innovation Program, College of Interdisciplinary Studies, Thammasat University, Bangkok, Thailand
[2]Department of Computer and Information Science, Faculty of Applied Science,
King Mongkut's University of Technology North Bangkok (KMUTNB), Bangkok, Thailand

Article Info

Article history:

Received Mar 23, 2026
Revised Sep 4, 2026
Accepted Sep 10, 2026

Keywords:

CatBoost
Explainable AI
Feature engineering
Gradient boosting
Insurance fraud detection

ABSTRACT

Insurance fraud detection remains difficult because fraudulent claims are rare, claim attributes are heterogeneous, and operational models must be accurate and explainable for audit-oriented decision support. This study proposes a validation-safe machine learning pipeline for three public datasets: insurance claims, fraud oracle, and healthcare provider fraud. The pipeline integrates domain-driven feature engineering, imbalance handling, CatBoost tabular learning, validation-only threshold selection, ablation analysis, hyperparameter sensitivity testing, and SHAP/LIME explainability. To reduce data-leakage risk, model selection and threshold tuning are performed only on training and validation partitions, while the test set is reserved for final evaluation. Unlike single-dataset studies, the experiments compare component-level effects across vehicle, automobile, and healthcare fraud settings and report F1, ROC-AUC, and PR-AUC. The best validation-safe test F1 scores are 0.724, 0.301, and 0.679, with corresponding PR-AUC scores of 0.498, 0.283, and 0.764. Results show that threshold calibration is essential under severe imbalance, whereas feature engineering, class weighting, and resampling are dataset-dependent. The contribution is a reproducible evaluation procedure that identifies which components transfer across insurance domains and which require dataset-specific validation. The explanations also support claims triage by highlighting influential attributes for global screening and local case review.

This is an open access article under the CC BY-SA license.

Corresponding Author:

Chiabwoot Ratanavilisagul
Department of Computer and Information Science, Faculty of Applied Science
King Mongkut's University of Technology North Bangkok (KMUTNB)
Bangkok, Thailand
Email: chaibwoot.r@sci.kmutnb.ac.th

1. INTRODUCTION

Insurance fraud is a persistent economic and social problem across vehicle, property, health, and life insurance sectors. The coalition against insurance fraud estimates that fraudulent claims cost the United States approximately USD 308 billion annually [1], while the Federal Bureau of Investigation reports that non-health insurance fraud alone adds USD 400 to USD 700 in annual costs for a typical household [2]. These losses motivate data-driven fraud detection systems that can prioritize suspicious claims before manual review. Traditional rule-based screening and manual investigation are useful for known fraud patterns but adapt poorly to evolving behavior [3]. Machine learning provides a complementary approach by learning nonlinear patterns from claims data [4]. In tabular fraud detection, gradient-boosting algorithms such as XGBoost [5], LightGBM [6], and CatBoost [7] are strong candidates because they handle nonlinear interactions and mixed feature spaces.

However, insurance fraud detection has four methodological challenges that are often treated separately: severe class imbalance, mixed numerical and categorical predictors, domain heterogeneity across insurance lines, and the need for interpretable evidence. SMOTE-style resampling [8], SHAP [9], and LIME [10] address parts of this problem, but they do not by themselves establish whether a full fraud-detection pipeline generalizes across datasets.

The research gap is therefore not simply whether a classifier can achieve high ROC-AUC on one dataset, but whether feature engineering, balancing, threshold selection, and explainability remain useful when the insurance domain and fraud prevalence change. A recent systematic review shows that automobile-insurance fraud studies rely on few public datasets and heterogeneous evaluation designs [11], while cost-sensitive fraud learning demonstrates that the best imbalance strategy depends on the target distribution and business loss structure [12], [13].

This paper makes five contributions. First, it reformulates the pipeline around a validation-safe protocol in which thresholds are selected only on validation data, and the test set is held out until final evaluation. Second, it compares CatBoost pipeline variants against classical, ensemble, boosting, and neural-network baselines. Third, it adds explicit SMOTE and borderline-SMOTE comparisons, PR-AUC reporting, confidence intervals, and Wilcoxon testing. Fourth, it uses ablation and sensitivity analysis to identify dataset-specific component effects. Fifth, it links global SHAP and local LIME explanations to practical claims-review workflows. To place the contribution in context, the revised manuscript explicitly compares ensemble learning [14], [15], neural and hybrid augmentation [16], healthcare fraud benchmarks [17], fraud-detection surveys [18], [19], general imbalance learning [20], [21], PR-AUC evaluation [22], nonparametric statistical comparison [23], and recent explainable healthcare fraud studies [24], [25].

2. LITERATURE REVIEW AND RESEARCH GAP

2.1. Machine learning for fraud detection

Fraud-detection surveys and analytics texts show that data mining can improve adaptive pattern recognition compared with fixed rules [3], [4]. Later reviews broadened this view by showing that fraud applications differ sharply by transaction type, label availability, cost structure, and evaluation metric [18], [19]. These studies are valuable, but much of the prior insurance literature remains descriptive or single-dataset, which limits claims about cross-domain robustness.

Tree ensembles remain strong for tabular fraud data because they model nonlinear interactions without requiring large sample sizes. Random forests reduce variance through bootstrap aggregation [14], while gradient boosting fits additive models that sequentially correct previous errors [15]. The limitation is that high ranking performance alone does not guarantee operational fraud recall or acceptable false-positive rates under rare-event conditions.

2.2. Gradient boosting algorithms

XGBoost [5] and LightGBM [6] provide efficient gradient boosting for structured data. CatBoost [7] is particularly suitable for insurance datasets because it reduces target leakage in categorical encoding through ordered boosting and performs well with limited hyperparameter tuning. This justifies CatBoost as the primary pipeline learner, while the experiments still compare it against alternative classifiers to avoid attributing all improvements to the model choice alone.

Deep learning and hybrid models are relevant but not automatically superior for these public tabular datasets. GAN-based augmentation has improved fraud classification in credit-card settings [16], and broader deep-learning imbalance studies show that network depth can amplify imbalance sensitivity when sample size and representation quality are limited [20], [21]. For this reason, the present study treats neural MLPs as baselines and focuses the proposed contribution on a transparent, reproducible tabular pipeline.

2.3. Handling class imbalance

Insurance fraud datasets exhibit severe class imbalance, so accuracy and ROC-AUC can be misleading when the minority fraud class is operationally important. SMOTE generates synthetic minority samples [8], whereas cost-sensitive learning changes the loss surface by penalizing minority-class errors more heavily [12], [13]. General imbalance research also recommends evaluating multiple strategies rather than assuming one universal remedy [20]. This study therefore compares raw CatBoost, class weighting, SMOTE, and Borderline-SMOTE under the same validation-safe thresholding protocol.

2.4. Explainable artificial intelligence (XAI)

Regulated insurance workflows require explanations that can be reviewed by analysts and auditors. SHAP provides additive global and local feature contributions [9], while LIME approximates local decision

boundaries around individual predictions [10]. These tools are useful only when paired with careful validation; otherwise, explanations may rationalize an overfit or leakage-prone model.

2.5. Cross-dataset and multi-domain evaluation

Cross-dataset evaluation is central to the present study. Schrijver *et al.* [11] found that automobile-insurance fraud research often depends on a small number of public datasets, and Schalck [12] showed that cost-sensitive and cost-insensitive fraud strategies can lead to different operating tradeoffs. Medicare fraud studies further show that provider-level signals differ from automobile or claim-level signals [17], [24], [25]. The novelty of this paper is to synthesize these strands into a single pipeline evaluation: each component is tested across three heterogeneous datasets, and the discussion explicitly separates model choice effects from pipeline-design effects.

3. METHOD

3.1. Datasets

- Dataset 1 — insurance claims (IC): 1,000 records, 40 features, binary target fraud_reported, fraud rate 24.7%. Features include policyholder demographics, vehicle characteristics, incident details, and claim amounts. This moderately imbalanced dataset provides a baseline multi-classifier evaluation.
- Dataset 2 — fraud oracle (FO): 15,420 automobile insurance claims, 33 features, binary target FraudFound_P, fraud rate 6.0%. This severely imbalanced dataset challenges standard threshold-based classifiers.
- Dataset 3 — healthcare provider fraud (HF): 5,410 provider-level aggregated records derived from CMS medicare claims, 27 features, binary target PotentialFraud, fraud rate 9.4%. Features capture claim volumes, reimbursement statistics, and patient diagnoses per provider.

Table 1 summarizes the three datasets used in the study, including sample size, feature count, fraud prevalence, application domain, and imbalance level. This summary is provided before the preprocessing description to clarify why the same validation-safe pipeline must be evaluated under different fraud-rate and domain conditions.

Table 1. Dataset summary

Dataset	Samples	Features	Fraud rate	Domain	Imbalance level
Insurance claims	1,000	40	24.7%	Vehicle	Moderate
Fraud oracle	15,420	33	6.0%	Auto	Severe
Health fraud provider	5,410	27	9.4%	Healthcare	Severe

3.2. Data processing

A standardized preprocessing pipeline is applied to all three datasets. Missing categorical values are imputed from the training data mode and continuous values from the training data median. Categorical variables are label-encoded for classical baselines, while CatBoost is used as the primary learner because its ordered boosting design reduces categorical target leakage [7]. Each dataset is split using stratified sampling into training, validation, and held-out test partitions (60%, 20%, and 20%). Scaling, imputation parameters, class-balancing operations, and feature thresholds are fitted only on training data. Validation data are used only for decision-threshold selection; the test set is not used for tuning.

Figure 1 illustrates the complete validation-safe workflow, from public datasets and stratified splitting through feature engineering, class-imbalance handling, CatBoost training, validation-only threshold tuning, and final held-out test evaluation. The figure clarifies that validation data are used for threshold selection and that the test set is reserved for final metrics only.

3.3. Feature engineering

- Insurance claims: six composite features — (i) claim_ratio: total claim/vehicle value; (ii) claim_diversity: distinct claim types per policyholder; (iii) injury_ratio: proportion of bodily injury claims; (iv) vehicle_ratio: vehicle age relative to claim amount; (v) tenure_group: policyholder tenure in risk-stratified bins; (vi) high_claim_flag: binary indicator for claims above 75th percentile.
- Fraud oracle: six features — (i) high_deductible; (ii) high_driver_risk; (iii) age_group; (iv) low_age_high_risk interaction; (v) policy_year_flag for first-year policies; (vi) deductible_risk_score.
- Healthcare provider fraud: six features — (i) claims_per_patient; (ii) reimbursement_per_claim; (iii) IP_OP_claims_ratio; (iv) physician_concentration (Herfindahl-hirschman index); (v) high_chronic_flag; (vi) deductible_reimb_ratio.

The engineered features are intentionally ratio-, frequency-, and risk-flag transformations rather than dataset identifiers. This design supports partial transferability: the exact variables differ by domain, but the underlying fraud concepts (claim intensity, reimbursement magnitude, provider concentration, policy risk, and abnormal claim ratios) are comparable. The ablation study tests whether these hand-crafted features add signal or introduce noise in each dataset.

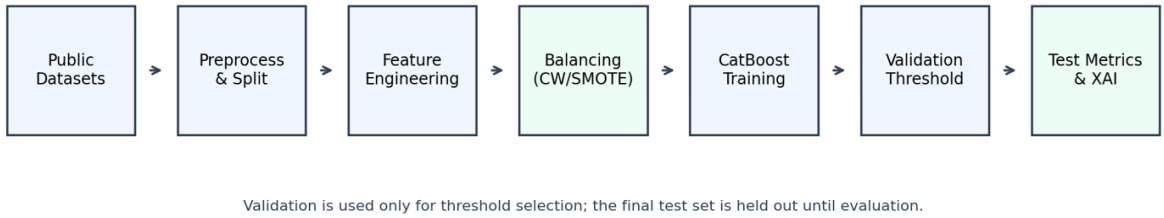

Figure 1. Proposed validation-safe machine learning pipeline

3.4. Pipeline configuration

Five configurations isolate each component's contribution:

- Config A: full pipeline (FE + CW + TT) — feature engineering, class-weight balancing, threshold tuning.
- Config B: no threshold tuning (FE + CW) — default 0.5 decision threshold.
- Config C: no class-weight (FE + TT) — feature engineering and threshold tuning only.
- Config D: no feature engineering (CW + TT) — original features with balancing and tuning.
- Config E: raw CatBoost — vanilla CatBoost, no component adjustments.

Class weights are set inversely proportional to class frequencies, with w_pos = n_negative/n_positive and w_neg = 1. For validation-threshold tuning, predicted probabilities are swept over thresholds from 0.05 to 0.95, and the threshold maximizing validation F1 is selected before final test evaluation. The F1-score is computed as F1 = 2 × Precision × Recall/(Precision + Recall). CatBoost minimizes binary log-loss, L = -1/N sum_i [y_i log(p_i) + (1-y_i) log(1-p_i)], with optional class weights applied to the minority class. ROC-AUC and PR-AUC are reported as complementary ranking metrics, with PR-AUC emphasized for severe class imbalance [22].

SMOTE and borderline-SMOTE are included as additional imbalance baselines. To avoid data leakage, synthetic samples are created only from the training partition or from the training fold inside cross-validation, never from validation or test data.

3.5. Ablation study design

A leave-one-out (LOO) ablation study on the full pipeline (Config A) individually removes each engineered feature and records the change in CV F1 and ROC-AUC. A separate cumulative addition experiment incrementally adds features to quantify marginal value. The ablation design isolates which features drive improvement and reveals potential redundancy.

3.6. Hyperparameter sensitivity analysis

Decision threshold: threshold swept from 0.05 to 0.90 in steps of 0.01; F1, precision, recall recorded. Boosting iterations: number of trees tested at {50, 100, 150, 200, 300, 400, 500} with 5-fold CV. Learning rate: tested at {0.005, 0.01, 0.02, 0.05, 0.1, 0.2, 0.3}; convergence and generalization were analyzed. Tree depth: maximum depth tested at {3, 4, 5, 6, 7, 8, 9, 10}; expressiveness vs. overfitting tradeoff characterized. All sensitivity analyses use 5-fold stratified cross-validation on the training set. The CatBoost reference configuration uses 300 iterations, a 0.05 learning rate, and depth 6.

For reproducibility, all experiments use random seed 42, stratified sampling, five-fold cross-validation for baseline and component comparisons, CatBoost 1.2.10 for the revised validation-safe run, and scikit-learn/imbalanced-learn for metrics and resampling. The generated CSV outputs, plots, and scripts are retained with the revised manuscript folder so that the reported tables can be traced to executable artifacts.

3.7. Explainability methods

Two complementary explainability methods are applied to the best-performing model per dataset. TreeSHAP [9] provides exact Shapley value computation for tree-based models; global importance is derived as mean(|SHAP_i|) over the test set, and beeswarm plots reveal both direction and magnitude of each feature's contribution. LIME [10] constructs a locally faithful linear approximation for individual fraud

predictions by perturbing the input instance, highlighting which feature values most strongly triggered the fraud classification in a specific claim.

4. RESULTS AND DISCUSSION

4.1. Pipeline configuration comparison

Table 2 presents validation-safe test performance for CatBoost pipeline configurations. Figure 2 summarizes the best F1 and PR-AUC achieved on each dataset. Because threshold tuning is now performed on validation data only, these values are more conservative than leakage-prone test-threshold estimates and better reflect expected deployment behavior.

Table 2. Pipeline configuration results (CatBoost, test set)

Dataset	Config.	Val. Thr.	F1	PR-AUC	ROC-AUC	Prec.	Recall
Insurance claims	A: Full (FE+CW+VT)	0.140	0.712	0.487	0.814	0.609	0.857
Insurance claims	B: No VT (FE+CW)	0.500	0.667	0.487	0.814	0.610	0.735
Insurance claims	C: No CW (FE+VT)	0.110	0.724	0.498	0.819	0.627	0.857
Insurance claims	D: No FE (CW+VT)	0.160	0.672	0.497	0.819	0.562	0.837
Insurance claims	E: Raw CatBoost	0.500	0.562	0.504	0.816	0.574	0.551
Fraud oracle	A: Full (FE+CW+VT)	0.530	0.283	0.240	0.838	0.187	0.589
Fraud oracle	B: No VT (FE+CW)	0.500	0.271	0.240	0.838	0.172	0.632
Fraud oracle	C: No CW (FE+VT)	0.120	0.301	0.283	0.851	0.208	0.541
Fraud oracle	D: No FE (CW+VT)	0.610	0.270	0.237	0.836	0.204	0.400
Fraud oracle	E: Raw CatBoost	0.500	0.063	0.263	0.853	1.000	0.032
Health fraud	A: Full (FE+CW+VT)	0.570	0.664	0.756	0.962	0.588	0.762
Health fraud	B: No VT (FE+CW)	0.500	0.651	0.756	0.962	0.547	0.802
Health fraud	C: No CW (FE+VT)	0.280	0.679	0.764	0.964	0.625	0.743
Health fraud	D: No FE (CW+VT)	0.630	0.658	0.747	0.957	0.597	0.733
Health fraud	E: Raw CatBoost	0.500	0.644	0.774	0.961	0.750	0.564

VT=validation-threshold tuning; FE=feature engineering; CW=class-weight; PR-AUC=area under the precision-recall curve

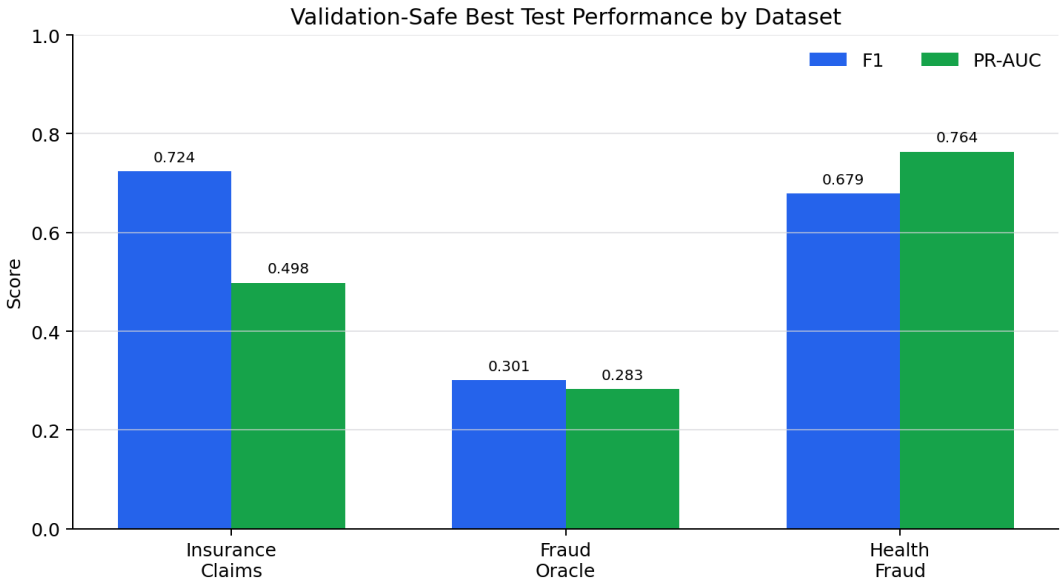

Figure 2. Best validation-safe test F1 and PR-AUC by dataset

The revised results show that pipeline components are dataset-dependent rather than universally additive. Insurance claims achieve its best validation-safe F1 with feature engineering and validation-threshold tuning without class weights (F1=0.724), while fraud oracle remains difficult despite high ROC-AUC because only 6.0% of cases are fraud. Healthcare provider fraud obtains the strongest PR-AUC (0.764) and best F1 of 0.679 without class weights, indicating that provider-level features already contain strong separability and that additional weighting can increase false positives.

4.2. Multi-classifier baseline comparison

Table 3 presents the strongest baseline models by dataset and includes PR-AUC to make the rare-class tradeoff explicit. The baseline comparison clarifies that performance gains cannot be attributed to CatBoost alone: LightGBM is competitive on insurance claims, Naive Bayes obtains higher default-threshold F1 on fraud oracle, and AdaBoost is competitive on healthcare provider fraud. CatBoost is therefore retained as the primary learner because of its robust tabular behavior and categorical-feature design, not because it dominates every metric in every dataset.

Table 3. Baseline model comparison – top models by CV F1 (No augmentation)

Dataset	Top model	CV F1	Test F1	ROC-AUC	PR-AUC	Prec.	Recall
Insurance claims	LightGBM	0.662	0.639	0.830	0.596	0.646	0.633
Insurance claims	CatBoost	0.635	0.641	0.826	0.537	0.611	0.673
Fraud oracle	Naive Bayes	0.159	0.180	0.772	0.138	0.152	0.222
Fraud oracle	CatBoost	0.077	0.062	0.858	0.272	0.750	0.032
Health fraud	CatBoost	0.595	0.659	0.963	0.775	0.756	0.584
Health fraud	AdaBoost	0.571	0.667	0.959	0.760	0.759	0.594

4.3. Class-imbalance handling comparison

Reviewer-requested imbalance baselines are shown in Table 4 and Figure 3. The comparison indicates that SMOTE and borderline-SMOTE improve PR-AUC on insurance claims but do not dominate fraud oracle or healthcare provider fraud. This supports the conclusion that imbalance handling must be selected empirically for each insurance domain rather than assumed to be universally beneficial.

Table 4. Validation-safe class-imbalance handling comparison

Dataset	Method	Val. Thr.	F1	PR-AUC	ROC-AUC	Prec.	Recall
Insurance claims	Raw	0.110	0.724	0.498	0.819	0.627	0.857
Insurance claims	Class-weight	0.140	0.712	0.487	0.814	0.609	0.857
Insurance claims	SMOTE	0.150	0.689	0.552	0.855	0.575	0.857
Insurance claims	Borderline-SMOTE	0.160	0.700	0.554	0.850	0.592	0.857
Fraud oracle	Raw	0.120	0.301	0.283	0.851	0.208	0.541
Fraud oracle	Class-weight	0.530	0.283	0.240	0.838	0.187	0.589
Fraud oracle	SMOTE	0.160	0.286	0.230	0.839	0.203	0.486
Fraud oracle	Borderline-SMOTE	0.140	0.288	0.218	0.841	0.191	0.584
Health fraud	Raw	0.280	0.679	0.764	0.964	0.625	0.743
Health fraud	Class-weight	0.570	0.664	0.756	0.962	0.588	0.762
Health fraud	SMOTE	0.520	0.639	0.748	0.957	0.545	0.772
Health fraud	Borderline-SMOTE	0.440	0.608	0.707	0.955	0.497	0.782

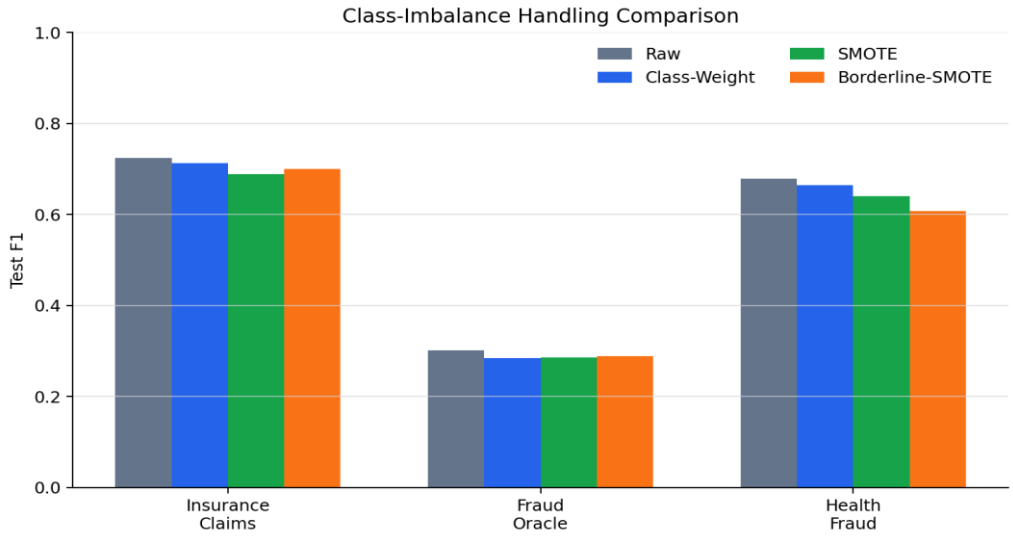

Figure 3. Test F1 comparison across imbalance-handling methods

4.4. Ablation study results

Table 5 summarizes the LOO ablation results. Feature names are generalized for cross-dataset comparison. Specific dataset features: IC uses claim_ratio, claim_diversity, injury_ratio, vehicle_ratio, tenure_group, high_claim_flag; FO uses high_deductible, high_driver_risk, age_group, low_age_high_risk, policy_year_flag, deductible_risk_score; HF uses claims_per_patient, reimbursement_per_claim, IP_OP_claims_ratio, physician_concentration, high_chronic_flag, deductible_reimb_ratio.

The ablation results reinforce the cross-dataset interpretation. For insurance claims, several engineered variables provide little incremental value, which is consistent with the revised validation-safe result where feature engineering helps only modestly. For healthcare provider fraud, dropping provider-intensity and concentration features reduces CV F1, confirming that these engineered features capture meaningful domain behavior. For fraud oracle, the small deltas indicate that the dominant challenge is not a missing handcrafted feature but a combination of severe imbalance and weak minority-class signal.

Table 5. LOO ablation study (CV F1 mean, CatBoost full pipeline)

Configuration	Ins. claims CV F1	Fraud oracle CV F1	Health fraud CV F1
All features (Full FE) — Baseline	0.704	0.290	0.609
Drop feature 1 (Rate/frequency)	0.713 (Δ+0.009)	0.283 (Δ-0.007)	0.601 (Δ-0.008)
Drop feature 2 (Diversity/volume)	0.715 (Δ+0.011)	0.287 (Δ-0.003)	0.606 (Δ-0.003)
Drop feature 3 (Injury/risk type)	0.706 (Δ+0.002)	0.290 (Δ+0.000)	0.618 (Δ+0.009)
Drop feature 4 (Ratio/concentration)	0.725 (Δ+0.021)	0.288 (Δ-0.002)	0.597 (Δ-0.012)
Drop feature 5 (Group/year flag)	0.716 (Δ+0.012)	0.282 (Δ-0.008)	0.604 (Δ-0.005)
Drop feature 6 (Binary risk flag)	0.708 (Δ+0.004)	0.288 (Δ-0.002)	0.600 (Δ-0.009)
No feature engineering (Original only)	0.708 (Δ+0.004)	0.285 (Δ-0.005)	0.600 (Δ-0.009)

4.5. Hyperparameter sensitivity analysis

Table 6 summarizes the optimal hyperparameter settings identified from four sensitivity analyses. The threshold analysis (S16a) reveals that the optimal decision threshold varies substantially: 0.24 for insurance claims (24.7% fraud), 0.69 for fraud oracle (6.0% fraud), and 0.80 for healthcare provider fraud (9.4% fraud). This inverse relationship between optimal threshold and fraud rate is intuitively consistent: lower fraud rates require the model to be more conservative (higher threshold) to avoid excessive false positives.

The learning rate analysis (S16c) shows consistent improvement as learning rate decreases from 0.3 to 0.005 across all datasets, suggesting the CatBoost default (0.05) is suboptimal for these datasets. The tree depth analysis (S16d) shows that shallow trees (depth 3) generalize better than deeper trees, consistent with classical variance-bias tradeoff analysis for boosted ensembles.

Table 6. Hyperparameter sensitivity summary (Best values by CV F1)

Experiment	Prameter	Ins. claims best	Fraud orable best	Health fraud best/observation
S16a	Threshold	0.24 (F1=0.730)	0.69 (F1=0.325)	0.80 (F1=0.690) — Optimal threshold shifts inversely with fraud rate
S16b	Iterations	50 (F1=0.730)	300 (F1=0.325)	400 (F1=0.690) — Health fraud benefits from more trees
S16c	Learning rate	0.005 (F1=0.737)	0.005 (F1=0.296)	0.005 (F1=0.711) — Lower LR consistently improves CV F1 across all datasets
S16d	Tree depth	Depth 3 (F1=0.730)	Depth 3 (F1=0.301)	Depth 3 (F1=0.694) — Shallow trees reduce overfitting

4.6. Critical interpretation and practical implications

4.6.1. Effectiveness of pipeline components

Results confirm that no single component universally dominates across all datasets. The revised validation-safe results show that class weighting and engineered features can improve recall, but they may also reduce precision when the original feature space is already informative. This is why the strongest configuration differs by dataset: insurance claims benefits from threshold tuning, fraud oracle requires careful rare-class calibration, and healthcare provider fraud is driven mainly by provider-level utilization and reimbursement features.

These findings align with cross-insurance evidence that fraud-detection strategies are domain-specific [11]. They also extend prior work by showing which pipeline components transfer and which must be revalidated. The practical implication is that insurers should treat the proposed pipeline as an evaluation procedure rather than a fixed recipe.

4.6.2. The class imbalance challenge

The fraud oracle dataset illustrates a critical failure mode of default classifiers under severe imbalance: CatBoost has strong ROC-AUC but very low default-threshold F1 because only a small fraction of fraud cases are predicted positive. PR-AUC and validation-threshold tuning reveal this weakness more directly than ROC-AUC alone, supporting the recommendation of precision-recall evaluation for imbalanced classification [22].

The SMOTE comparison further shows that resampling is not automatically superior. SMOTE improves insurance claims PR-AUC but provides limited gains on fraud oracle and healthcare provider fraud. This pattern is consistent with imbalance-learning literature: synthetic sampling is useful only when minority neighborhoods are informative rather than noisy [8], [20].

4.6.3. Hyperparameter sensitivity insights

The threshold sensitivity curves reveal sharp F1 peaks at dataset-specific optima, confirming that threshold miscalibration has large operational consequences. Lower learning rates and shallower trees generally improve cross-validation F1, suggesting that conservative boosting reduces variance in small-to-medium imbalanced insurance datasets. However, these hyperparameters should be tuned within validation folds rather than selected from final test behavior.

4.6.4. Explainability analysis

SHAP analysis identifies claims_per_patient and reimbursement_per_claim as important healthcare provider fraud features, which is consistent with provider overbilling patterns reported in healthcare fraud studies [17], [24], [25]. For insurance claims, high_claim_flag and claim_ratio rank among the top explanatory features, confirming that unusually high claim amounts relative to policy or vehicle context are key fraud indicators. LIME local explanations complement SHAP by showing why an individual claim or provider was flagged. In practice, this gives claims adjusters a concise explanation for escalation while preserving the final decision for human review. The XAI outputs should therefore be interpreted as decision-support evidence, not as automated proof of fraud.

4.6.5. Limitations and future work

Several limitations should be noted. First, the public datasets are static and may not reflect changing fraud schemes. Second, some datasets are small or highly imbalanced, limiting statistical power; Wilcoxon tests comparing full and raw CatBoost show directional improvements for insurance claims and fraud oracle but do not reach $p < 0.05$ with five folds. Third, feature engineering remains domain-informed and should be revalidated before deployment. Fourth, neural, graph-based, and temporal models may capture higher-order relationships that are outside the scope of the present tabular pipeline.

Future work should evaluate temporal claim histories, graph neural networks for provider and claimant networks, cost-sensitive business loss functions, Bayesian hyperparameter optimization, and federated learning for privacy-preserving collaboration among insurers.

5. CONCLUSION

This paper presented a validation-safe, multi-dataset machine learning pipeline for insurance fraud detection across vehicle, automobile, and healthcare domains. The revised experiments show best test F1-scores of 0.724, 0.301, and 0.679 and best PR-AUC scores of 0.498, 0.283, and 0.764 for insurance claims, fraud oracle, and healthcare provider fraud, respectively.

The main finding is that pipeline components are not universally beneficial. Threshold selection is essential under imbalance, but feature engineering, class weighting, SMOTE, and borderline-SMOTE must be validated per dataset. This directly addresses the practical problem faced by insurers: a model with high ROC-AUC may still fail to identify enough fraud cases unless thresholds and rare-class metrics are handled explicitly.

Practically, the pipeline can support claims triage by ranking suspicious cases, exposing influential risk factors through SHAP and LIME, and reducing manual review burden while keeping final decisions auditable. Future work should extend the pipeline with temporal, graph-based, and privacy-preserving learning methods.

ACKNOWLEDGMENTS

The authors thank the anonymous reviewers for constructive comments that improved the structure, reference quality, dataset-bias discussion, and interpretation of the results.

FUNDING INFORMATION

This research received no specific research grant or contract from any funding agency in the public, commercial, or not-for-profit sectors.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Somkiat Kosolsombat	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓		✓	✓
Chiabwoot Ratanavilisagul	✓	✓		✓		✓	✓	✓		✓	✓	✓	✓	✓

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. The authors state no conflict of interest.

DATA AVAILABILITY

The authors confirm that the datasets used in this study are public or derived from public sources as cited in the manuscript. The revised experimental scripts, generated CSV files, and plots are available from the corresponding author upon reasonable request and can be deposited in a public repository if required by the journal.

REFERENCES

[1] Coalition Against Insurance Fraud, *The Impact of Insurance Fraud on the U.S. Economy*. Washington, DC, USA: Coalition Against Insurance Fraud, 2022. [Online]. Available: https://insurancefraud.org/wp-content/uploads/The-Impact-of-Insurance-Fraud-on-the-U.S.-Economy-Report-2022-8.26.2022.pdf

[2] Federal Bureau of Investigation, U.S. Department of Justice, *Financial Crimes Report to the Public: 2005*. Washington, DC, USA: Federal Bureau of Investigation, May 2005. [Online]. Available: FBI file repository. Accessed: May 29, 2026.

[3] K. G. Al-Hashedi and P. Magalingam, "Financial fraud detection applying data mining techniques: a comprehensive review from 2009 to 2019," *Computer Science Review*, vol. 40, p. 100402, May 2021, doi: 10.1016/j.cosrev.2021.100402.

[4] B. Baesens, V. Van Vlasselaer, and W. Verbeke, *Fraud Analytics Using Descriptive, Predictive, and Social Network Techniques*. Wiley, 2015. doi: 10.1002/9781119146841.

[5] T. Chen and C. Guestrin, "XGBoost: a scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA: ACM, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[6] G. Ke *et al.*, "LightGBM: a highly efficient gradient boosting decision tree," *Advances in Neural Information Processing Systems*, vol. 2017-December, pp. 3147–3155, 2017.

[7] P. Liudmila, G. Gleb, V. Aleksandr, D. Anna Veronika, and G. Andrey, "Catboost: unbiased boosting with categorical features," *Advances in Neural Information Processing Systems*, no. Section 4, pp. 6638–6648, 2018.

[8] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.

[9] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems*, pp. 4768–4777, 2017.

[10] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should i trust you?' Explaining the predictions of any classifier," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA: ACM, Aug. 2016, pp. 1135–1144. doi: 10.1145/2939672.2939778.

[11] G. Schrijver, D. K. Sarmah, and M. El-hajj, "Automobile insurance fraud detection using data mining: a systematic literature review," *Intelligent Systems with Applications*, vol. 21, p. 200340, Mar. 2024, doi: 10.1016/j.iswa.2024.200340.

[12] M. Y.Schalck, "Auto insurance fraud detection: leveraging cost sensitive and insensitive algorithms for comprehensive analysis," *Insurance: Mathematics and Economics*, vol. 122, pp. 44–60, May 2025, doi: 10.1016/j.insmatheco.2025.02.001.

[13] Y. Sun, M. S. Kamel, A. K. C. Wong, and Y. Wang, "Cost-sensitive boosting for classification of imbalanced data," *Pattern Recognition*, vol. 40, no. 12, pp. 3358–3378, Dec. 2007, doi: 10.1016/j.patcog.2007.04.009.

[14] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[15] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, Oct. 2001, doi: 10.1214/aos/1013203451.
[16] U. Fiore, A. De Santis, F. Perla, P. Zanetti, and F. Palmieri, "Using generative adversarial networks for improving classification effectiveness in credit card fraud detection," *Information Sciences*, vol. 479, pp. 448–455, Apr. 2019, doi: 10.1016/j.ins.2017.12.030.
[17] R. A. Bauder and T. M. Khoshgoftaar, "Medicare fraud detection using machine learning methods," in *Proceedings - 16th IEEE International Conference on Machine Learning and Applications, ICMLA 2017*, IEEE, Dec. 2017, pp. 858–865. doi: 10.1109/ICMLA.2017.00-48.
[18] A. Abdallah, M. A. Maarof, and A. Zainal, "Fraud detection system: a survey," *Journal of Network and Computer Applications*, vol. 68, pp. 90–113, Jun. 2016, doi: 10.1016/j.jnca.2016.04.007.
[19] 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, Feb. 2011, doi: 10.1016/j.dss.2010.08.006.
[20] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009, doi: 10.1109/TKDE.2008.239.
[21] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *Journal of Big Data*, vol. 6, no. 1, p. 27, Dec. 2019, doi: 10.1186/s40537-019-0192-5.
[22] 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, p. e0118432, Mar. 2015, doi: 10.1371/journal.pone.0118432.
[23] D. Janez, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine Learning Research*, vol. 7, no. 1, pp. 1–30, 2006.
[24] S. Shekhar, J. Leder-Luis, and L. Akoglu, "Unsupervised machine learning for explainable health care fraud detection," Cambridge, MA, Feb. 2023. doi: 10.3386/w30946.
[25] E. Nabrawi and A. Alanazi, "Fraud detection in healthcare insurance claims using machine learning," *Risks*, vol. 11, no. 9, p. 160, Sep. 2023, doi: 10.3390/risks11090160.

BIOGRAPHIES OF AUTHORS



Somkiat Kosolsombat received the bachelor's degree in Computer Science from Mahidol University, Thailand, in 1995, and the master's degree in Computer Science from Thammasat University, Thailand, in 2011, and Ph.D. in Computer Science from King Mongkut's University of Technology North Bangkok (KMUTNB), Thailand, in 2023. He is currently a full-time lecturer at the Department of Data Science and Innovation, College of Interdisciplinary Studies, Thammasat University, Thailand. His research interests include artificial intelligence, machine learning, big data, and information retrieval. He can be contacted at email: somkiatk@tu.ac.th.



Chiabwoot Ratanavilisagul received the bachelor's degree in Computer Engineering from King Mongkut's Institute of Technology Ladkrabang (KMITL), Thailand, in 2010, the master's degree in Computer Engineering from KMITL in 2013, and the Doctor's degree in Electrical Engineering from KMITL in 2015. He is currently a full-time lecturer at the Department of Computer and Information Science, Faculty of Applied Science, King Mongkut's University of Technology North Bangkok, Thailand. His research interests include artificial intelligence, natural language processing, machine learning, and optimization. He can be contacted at email: chaibwoot@hotmail.com.