

Model-Dependent Calibration Effects of Imbalance Handling in Credit Card Fraud Detection

Itishom Al Khoiry^{1*}, Ika Menarianti², Shofif Sobaruddin Akbar³, Ratih Hesty Utami Puspitasari⁴

^{1,2}Department of Digital Business, Universitas Persatuan Guru Republik Indonesia Semarang, Indonesia

^{3,4}Department of Management, Universitas Persatuan Guru Republik Indonesia Semarang, Indonesia

DOI: <https://doi.org/10.51583/IJLTEMAS.2026.150900166>

Received: 07 October 2026; Accepted: 12 October 2026; Published: 26 October 2026

ABSTRACT

Class imbalance correction methods such as class weighting and synthetic oversampling are widely used to improve fraud detection, but they can distort the probability estimates that risk scoring depends on. This study evaluates the calibration cost of two common imbalance handling strategies, inverse frequency class weighting and SMOTE-NC, applied to three model families: Logistic Regression, Random Forest, and XGBoost. Using the Sparkov credit card transaction dataset, comprising 1,296,675 training transactions and 555,719 future test transactions, the analysis follows a chronological four way split that separates model training, calibration fitting, threshold selection, and future testing, so that no evaluation period leaks information into model fitting or calibration. Across the three models, mean raw expected calibration error (ECE) rose from 0.0012 without imbalance handling to 0.0715 with class weighting and 0.0742 with SMOTE-NC, driven mainly by Logistic Regression. Mean raw F1 score declined from 0.680 to 0.648 and 0.622, respectively, showing that neither strategy consistently improved detection performance. The effects were model dependent: class weighting improved the F1 score of XGBoost from 0.7878 to 0.8106 while substantially increasing its calibration error, whereas Random Forest with class weighting retained near ideal calibration. For the five configurations whose raw ECE exceeded 0.01, post hoc isotonic calibration, fitted on a dedicated chronological period, reduced ECE by 96.6% to 99.4% while keeping future test F1 score differences below 0.008 across all nine raw versus calibrated pairs. The calibration gains were robust across multiple ECE binning schemes, including 10, 20, and 50 equal width bins and adaptive equal frequency binning. These results indicate that imbalance handling should be evaluated separately for each model family rather than assumed to help uniformly, and that post hoc calibration should be applied conditionally, only when diagnostics indicate substantial probability distortion, rather than as a default step in every fraud detection pipeline.

Keywords: credit card fraud detection, probability calibration, class imbalance, SMOTE-NC, class weighting

INTRODUCTION

Digital financial transactions have grown rapidly, and with them the volume and sophistication of fraudulent activity. Machine learning is widely used for fraud detection because it can identify complex patterns in historical transactions that fixed rules may miss [1], [2]. Fraud data, however, are extremely imbalanced, with genuine transactions often outnumbering fraudulent ones by several hundred to one [3], [4]. A model trained without accounting for this imbalance may achieve high accuracy by favoring the majority class while still missing the rare fraud cases of interest [5].

To counter this, practitioners routinely apply imbalance-handling techniques. The two most common families are oversampling, of which SMOTE and its mixed-type variant SMOTE-NC are the best known [6], and class weighting, which increases the loss contribution of the minority class. While these approaches can improve the detection of minority-class observations, they also modify the learning conditions under which the model

estimates class probabilities. This may affect not only classification performance but also the reliability of the resulting probability estimates. These techniques are widely reported to raise recall on fraud benchmarks [7]–[9], and are now applied almost by default.

A question that is far less often asked is whether the probabilities these models produce remain trustworthy. It is useful to separate two distinct uses of a model score. For ranking, where analysts review the top-scoring transactions under a fixed budget, the score order is what matters and calibration is not required. For absolute-

threshold decisions, where an action is triggered when the estimated fraud probability exceeds a fixed value, or where the score is reported as a probability, the numeric value must be reliable [10], [11]. It is this second, interpretive use that calibration governs. Calibration and discrimination are different properties [12], [13]. Resampling in particular shifts the base rate seen during training away from the true operating base rate, which can bias probability estimates toward the minority class [14], [15]. Despite this, most fraud-detection studies report only threshold-based detection metrics and omit any calibration assessment. This omission leaves an important gap. The calibration effects of class weighting and SMOTE-NC may vary across model families. It is therefore necessary to examine whether isotonic calibration can improve probability reliability while maintaining detection performance under independently selected thresholds.

This study measures how imbalance handling affects probability calibration in fraud detection and, equally important, tests a corrective procedure. The contributions are threefold. First, we quantify the calibration cost of class weighting and SMOTE-NC under a chronological four-way design that separates model training, calibration fitting, threshold selection, and future testing, so that neither the calibrator nor the operating threshold borrows information from the evaluation period. Second, we examine whether the calibration effects vary across model families and are accompanied by changes in classification performance. Third, we evaluate post hoc isotonic calibration as a corrective step and report both where it works and where it does not, and we derive practical guidance for the conditional use of post hoc calibration that practitioners can apply directly.

RELATED WORK

Machine Learning for Fraud Detection

Supervised learning is widely used in fraud detection [1], [2]. Tree ensembles, including Random Forest and gradient boosting models such as XGBoost, have performed well on tabular fraud data [7], [8], [16]. Their performance is commonly evaluated using accuracy, precision, recall, F1 score, and ROC AUC [9]. However, these metrics primarily assess discrimination and classification performance rather than the reliability of predicted probabilities. Probability calibration offers a complementary perspective by examining whether predicted risk scores correspond to the fraud frequencies observed in practice [10], [11], [17]. This distinction is particularly important when model outputs are used as risk probabilities rather than only for ranking transactions.

Handling Class Imbalance

There are two approaches to dealing with class imbalance: data level and algorithm level [3], [5]. Data level approaches involve altering the training distribution by resampling, such as SMOTE and its variants [18]. SMOTE-NC accommodates mixed numeric and categorical attributes by preserving valid categorical values when generating synthetic observations [6]. The learning objective is changed without altering the data distribution in algorithm level approaches such as class weighting and cost sensitive learning [19]. Resampling can also alter the distribution of predicted scores [14]. Dal Pozzolo et al. demonstrate that changes in the training distribution can shift posterior probabilities and may require probability correction to reflect the original class distribution [15].

Probability Calibration

A classifier is calibrated when its predicted probabilities correspond to observed frequencies [20], [21]. Calibration can be assessed using ECE, the Brier score, and reliability diagrams [22], [23]. Post hoc methods

adjust model scores using separate calibration data, with Platt scaling applying a sigmoid function and isotonic regression applying a monotonic mapping [24], [25]. Calibration under class imbalance has been studied in the context of general classification and medical imaging [26], [27]. Liu [28] compares SMOTE and undersampling on five datasets and two ensemble families, finding that SMOTE results in an ECE increase of 0.009, while Lin et al. [29] examine probability estimation under imbalanced conditions. The present study builds on this literature and examines the use of SMOTE-NC and class weighting in fraud risk scoring, and uses a chronological four way design to separate calibration fitting, threshold selection, and future testing.

METHODOLOGY

Dataset

The analysis draws on the Sparkov credit card transaction dataset [30], covering January 2019 to December 2020. It comprises 1,296,675 training transactions with 7,506 fraud cases (0.579%) and 555,719 test transactions with 2,145 fraud cases (0.386%), indicating a substantial class imbalance in both subsets. As shown in Fig. 1, fraudulent transactions tend to involve higher amounts, while fraud rates vary considerably across merchant categories, highlighting the potential relevance of these features for fraud detection.

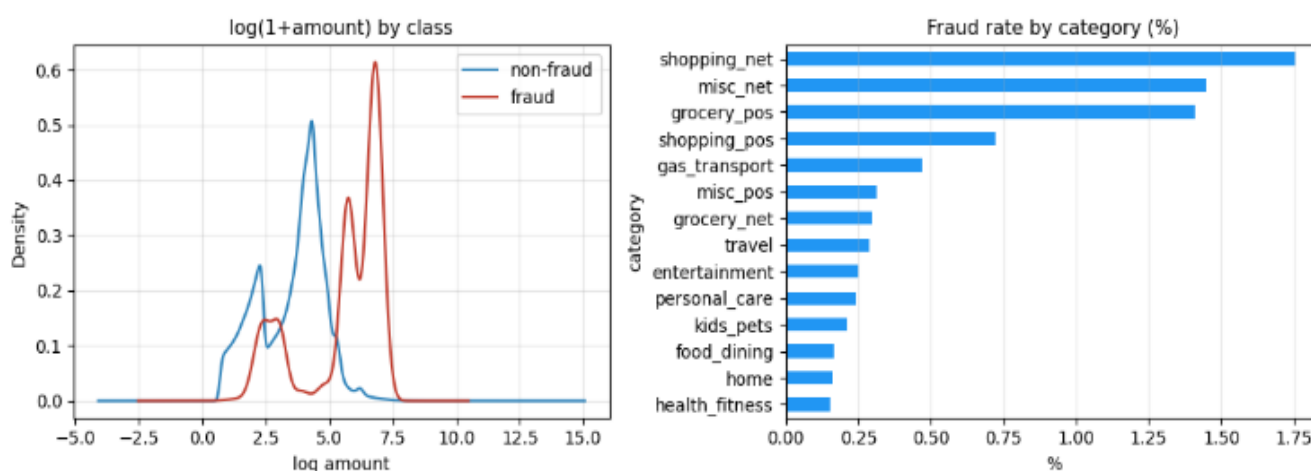


Figure 1. Distribution of log-transformed transaction amount by class (left) and fraud rate by merchant category (right). Fraud concentrates at higher amounts and in online categories.

Feature Engineering

Direct identifiers, including card number, name, street, and transaction identifier, are removed to prevent the model from memorizing record identities. The derived features include transaction hour, day of week, weekend status, cardholder age, distance between the cardholder and merchant calculated using the haversine formula, and log-transformed transaction amount. Merchant category and gender are encoded using one-hot encoding, while numeric features are standardized using statistics from the training period, producing 28 model inputs.

Chronological Four-Way Split

To prevent future information leakage, all transactions are ordered chronologically and divided into four disjoint periods, as summarized in Table 1. These periods are used separately for model training, isotonic calibration, F1 maximizing threshold selection, and future testing. Separate thresholds are selected for raw and calibrated probabilities, then fixed before evaluation on the untouched future test period. This design ensures that the test data do not inform model fitting, calibration, or threshold selection. Fraud prevalence declines from 0.5753% during training to 0.3860% during future testing, allowing evaluation under a change in prevalence over time.

Table 1. Chronological Four-Way Split

Period	From	To	Records	Fraud	Rate	Role
Training	1/1/2019	3/6/2020	1,037,340	5,968	0.58%	model fitting
Calib. fit	3/7/2020	5/2/2020	129,667	706	0.54%	calibrator
Thr. sel.	5/3/2020	6/21/2020	129,668	832	0.64%	threshold
Future test	6/22/2020	12/31/2020	555,719	2,145	0.39%	evaluation

Imbalance Handling and Models

We compare three imbalance strategies: no handling, inverse frequency class weighting, and SMOTE-NC. Class weighting modifies the training objective by assigning greater importance to minority cases, whereas SMOTE-NC generates additional minority observations. SMOTE-NC is applied only to the training period and before one-hot encoding to preserve valid merchant category and gender values. All evaluation periods retain their original distributions, ensuring that performance is assessed under the observed class imbalance. Logistic Regression serves as a linear diagnostic baseline, while Random Forest and XGBoost represent bagging and boosting models, allowing the effects of imbalance handling to be examined across different learning mechanisms.

Post Hoc Calibration and Threshold Selection

Isotonic regression is fitted on the calibration-fit period for each model-imbalance configuration to transform raw probabilities into calibrated probabilities. This method is chosen for its flexibility, as miscalibration introduced by resampling may not follow a sigmoidal pattern. Separate thresholds for raw and calibrated probabilities are then selected on the threshold-selection period to maximize the F1 score, balancing precision and recall. Each threshold is frozen before evaluation on the future-test period to prevent information leakage. Hyperparameters are predefined to limit the influence of hyperparameter search and support consistent comparisons across imbalance strategies, calibration conditions, and model families.

Evaluation Metrics and Reproducibility

Detection performance is evaluated using recall, precision, F1 score, and PR AUC, while calibration is assessed using ECE, Brier score, log loss, and reliability diagrams [5], [22], [23]. The robustness of ECE is investigated with 10, 20 and 50 equal width bins and adaptive equal frequency ECE. The three models, three imbalance strategies, and two probability conditions produce 18 configurations. Logistic Regression uses L2 regularization with 1,000 maximum iterations; Random Forest uses 200 trees; and XGBoost uses 300 trees, maximum depth 6, and learning rate 0.1. Class weighting uses inverse frequency weights, with `scale_pos_weight` set to the class ratio for XGBoost. SMOTE-NC balances the training data to 1,031,372 observations per class. All stochastic procedures use random seed 42, and isotonic calibration is fitted in prefit mode.

RESULTS AND DISCUSSION

Calibration Cost of Imbalance Handling

The mean raw results for the three models are summarized in Table 2. With class weighting, mean ECE rose from 0.0012 to 0.0715, while SMOTE-NC increased it to 0.0742. These ECE values represent weighted average calibration gaps of 7.15 and 7.42 percentage points, not classification error rates. Mean F1 score declined from 0.6801 to 0.6483 and 0.6223, respectively. However, these averages mask model differences: calibration deteriorated sharply for Logistic Regression, remained reliable for Random Forest with class weighting, and worsened for XGBoost with class weighting despite its improved F1 score, revealing a model-specific trade-off between calibration and classification performance.

Table 2. Mean Effect of Imbalance Strategy on Raw Probabilities (Averaged over Three Models)

Imbalance	Mean ECE	Mean Brier	Mean F1
None	0.0012	0.0018	0.6801
Class weight	0.0715	0.0278	0.6483
SMOTE-NC	0.0742	0.0274	0.6223

Model-Dependent Effects

Table 3 reports the configuration-level results, showing that the effects of imbalance handling differ across models. Among the raw configurations, Random Forest without handling achieved the highest F1 score of 0.8409, while class weighting produced a slightly higher PR AUC of 0.8732 and remained well calibrated with an ECE of 0.0005. For Random Forest, SMOTE-NC reduced F1 score and increased calibration error. XGBoost was the only model whose raw F1 score improved with class weighting, rising from 0.7878 to 0.8106, although recall declined from 0.7795 to 0.7650 and ECE increased from 0.000384 to 0.01114. For Logistic Regression, both class weighting and SMOTE-NC increased calibration error without improving F1 score. These results demonstrate that the effects of imbalance handling vary according to the model and the performance measure considered.

Table 3. Full Experimental Grid: Detection and Calibration Metrics on the Future Test Period

Model	Imbalance	Prob.	Recall	Precision	F1	PR-AUC	ECE	Brier
LR	None	Raw	0.3869	0.4399	0.4117	0.2921	0.0023	0.0031
LR	None	Isotonic	0.3776	0.4712	0.4193	0.2914	0.0012	0.003
LR	Class weight	Raw	0.676	0.1966	0.3046	0.1403	0.2028	0.0771
LR	Class weight	Isotonic	0.676	0.1967	0.3048	0.1627	0.0012	0.0034
LR	SMOTE-NC	Raw	0.6242	0.2367	0.3432	0.1632	0.1891	0.0719
LR	SMOTE-NC	Isotonic	0.6242	0.2365	0.3431	0.1938	0.0012	0.0033
RF	None	Raw	0.7823	0.909	0.8409	0.8701	0.0009	0.001
RF	None	Isotonic	0.7804	0.9118	0.841	0.8617	0.0002	0.001
RF	Class weight	Raw	0.8	0.8619	0.8298	0.8732	0.0005	0.001
RF	Class weight	Isotonic	0.7967	0.8728	0.833	0.8616	0.0002	0.001
RF	SMOTE-NC	Raw	0.71	0.8268	0.764	0.795	0.0115	0.003
RF	SMOTE-NC	Isotonic	0.7319	0.7929	0.7612	0.7812	0.0004	0.0014
XGB	None	Raw	0.7795	0.7962	0.7878	0.8488	0.0004	0.0012
XGB	None	Isotonic	0.7758	0.8062	0.7907	0.8361	0.0002	0.0012
XGB	Class weight	Raw	0.765	0.8619	0.8106	0.8519	0.0111	0.0052
XGB	Class weight	Isotonic	0.7716	0.8444	0.8063	0.8312	0.0003	0.0012
XGB	SMOTE-NC	Raw	0.7534	0.7659	0.7596	0.8058	0.0219	0.0072
XGB	SMOTE-NC	Isotonic	0.7478	0.7798	0.7634	0.7916	0.0003	0.0013

Effectiveness of the Corrective Procedure

Table 4 summarizes the effect of isotonic calibration across all model and imbalance configurations. Five configurations using imbalance handling had raw ECE above 0.01 on the original scale. This cutoff is used only as a descriptive grouping in this experiment and is not proposed as a universal calibration threshold. Isotonic calibration reduced ECE in these configurations by 96.6% to 99.4%. With separate thresholds selected for raw and calibrated probabilities on the independent threshold selection period, absolute F1 score differences on the future test period remained below 0.008 across all nine pairs. Thus, isotonic calibration

substantially reduced measured calibration error while maintaining comparable threshold based detection performance under separately selected operating thresholds.

Table 4. Effect of Post Hoc Isotonic Calibration per Configuration

Model / Imb.	ECE raw ($\times 10^{-3}$)	ECE isot. ($\times 10^{-3}$)	Reduction	F1 raw	F1 isot.
LR / CW	202.8	1.2	99.40%	0.3046	0.3048
LR / SMOTE-NC	189.1	1.2	99.40%	0.3432	0.3431
LR / none	2.3	1.2	49.60%	0.4117	0.4193
RF / SMOTE-NC	11.5	0.4	96.60%	0.764	0.7612
RF / CW	0.5	0.2	48.60%	0.8298	0.833
RF / none	0.9	0.2	71.90%	0.8409	0.841
XGB / CW	11.1	0.3	97.30%	0.8106	0.8063
XGB / SMOTE-NC	21.9	0.3	98.50%	0.7596	0.7634
XGB / none	0.4	0.2	53.30%	0.7878	0.7907

Note: ECE values are reported in units of $\times 10^{-3}$; reduction and F1-score in original scales.

Robustness of the Calibration Measurements

Table 5 evaluates whether the calibration findings depend on the ECE binning scheme. Across the selected Random Forest and XGBoost configurations, the results remain consistent using 10, 20, and 50 equal width bins and adaptive equal frequency binning. For Random Forest with SMOTE-NC and XGBoost with class weighting, raw ECE values remain approximately 0.0115 and 0.0111, respectively, across the different schemes and decrease to below 0.0004 after isotonic calibration. Log loss also improves in both configurations. These results indicate that the observed calibration patterns for the selected ensemble configurations are unlikely to be artifacts of a particular binning scheme.

Table 5. Calibration Robustness Across ECE Binning Schemes and Log Loss

Configuration	ECE-10	ECE-20	ECE-50	Adapt.	Log loss
RF none / raw	0.87	0.88	0.91	0.62	7.52
RF none / isot.	0.24	0.26	0.26	0.33	5.11
RF SM-NC / raw	11.54	11.54	11.54	11.56	17.35
RF SM-NC / isot.	0.39	0.39	0.39	0.5	7.53
XGB none / raw	0.38	0.38	0.43	0.05	5.45
XGB none / isot.	0.18	0.24	0.24	0.25	6.14
XGB CW / raw	11.11	11.11	11.11	11.11	19.48
XGB CW / isot.	0.3	0.32	0.37	0.34	5.55

Note: All ECE and log-loss values are reported in units of $\times 10^{-3}$.

Reliability Diagrams

XGBoost was selected for Figure 2 because it most clearly illustrates the tradeoff between detection performance and probability calibration: class weighting improved its F1 score but substantially increased ECE. Under class weighting and SMOTE-NC, the raw curves depart from the diagonal, indicating systematically overstated fraud risk, while isotonic calibration brings them closer to the diagonal, improving agreement between predicted and observed probabilities. Without imbalance handling, the raw and calibrated curves already lie near the diagonal, suggesting that the observed miscalibration was primarily associated with imbalance handling.

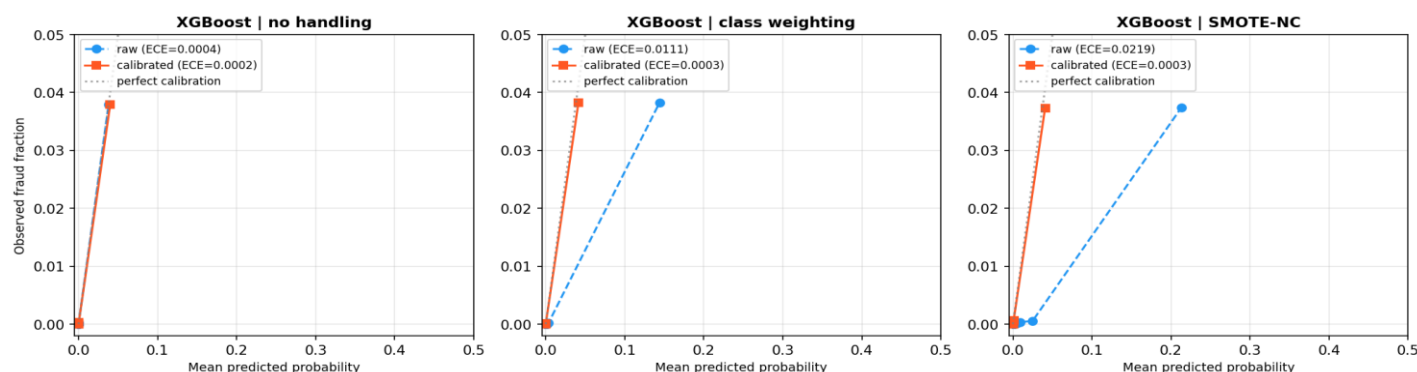


Figure 2. XGBoost reliability diagrams under no handling, class weighting, and SMOTE-NC. Isotonic calibration (squares) improves alignment with observed fraud rates.

Calibration Tradeoffs

Post hoc calibration should not be used in a blanket fashion. ECE dropped from 0.000384 to 0.000180 after isotonic calibration, while log loss rose from 0.005452 to 0.006142, the Brier score was almost the same, and PR AUC decreased from 0.8488 to 0.8361 for XGBoost without imbalance handling. This finding indicates that a lower ECE does not necessarily improve other probability or ranking metrics, especially when the raw model is already well calibrated. Therefore, decisions on calibration should be made based on ECE in combination with Brier score, log loss, and PR AUC, not just one of these metrics. This distinction matters because probability estimation and transaction ranking serve different operational purposes.

Practical Guidance for Conditional Calibration

The observed results support the conditional use of calibration. Raw probabilities should be assessed first with ECE, Brier score, log loss and a reliability diagram. If these diagnostics consistently indicate miscalibration, isotonic calibration can be fitted on a dedicated chronological calibration period, followed by threshold selection on a separate validation period. The calibrated pipeline should then be evaluated on future data, and raw probabilities should be retained when calibration provides no broader improvement. Calibration is most useful when scores are interpreted as probabilities or when absolute risk thresholds are used, but may not be necessary when scores are used for ranking tasks.

LIMITATIONS

This study provides a controlled evaluation of calibration effects across model families using the Sparkov benchmark and a single random seed. Future studies can examine the stability of these patterns using operational transaction data, repeated experiments with multiple random seeds, tuned configurations, and alternative calibration methods. Prospective evaluation with a prespecified calibration criterion can validate the proposed conditional workflow.

CONCLUSION

This study examined how class imbalance handling affects probability calibration in fraud detection and evaluated isotonic calibration as a corrective procedure. Averaged across the three models, class weighting and SMOTE-NC raised mean raw ECE from 0.0012 without imbalance handling to 0.0715 and 0.0742, respectively, although the increase was driven largely by Logistic Regression. Neither strategy consistently improved detection performance, and XGBoost achieved a higher F1 score with class weighting, demonstrating that the effects were model dependent. For the five configurations with raw ECE above 0.01, isotonic calibration reduced ECE by 96.6% to 99.4%. Across all nine raw and calibrated pairs, absolute future test F1 score differences remained below 0.008 when separate operating thresholds were selected. However, calibration provided limited or mixed benefits for models that were already well calibrated and did not consistently improve log loss or PR AUC. These findings demonstrate that imbalance handling and probability calibration should be evaluated jointly for each model. Post hoc calibration should be applied only

when diagnostics indicate substantial probability distortion, using dedicated chronological calibration data and separate threshold selection. Future research should validate this workflow on operational transaction data and under temporal distribution shifts.

Ethical Considerations

Ethical Approval

This study did not involve human participants, human tissue, or animal subjects. The analysis relies exclusively on the Sparkov dataset, a publicly available, simulated credit card transaction dataset released for fraud-detection research on Kaggle. No ethical approval was therefore required for this study.

Conflict of Interest

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.

Data Availability

The dataset analyzed in this study, the Sparkov credit card transaction dataset, is publicly available on Kaggle at <https://www.kaggle.com/datasets/kartik2112/fraud-detection>. No new data were collected for this study.

ACKNOWLEDGMENT

The authors gratefully acknowledge the academic and institutional support provided by Universitas Persatuan Guru Republik Indonesia Semarang, Indonesia.

REFERENCES

1. K. G. Dastidar, O. Caelen, and M. Granitzer, "Machine learning methods for credit card fraud detection: A survey," *IEEE Access*, vol. 12, pp. 158939–158965, 2024.
2. A. Cherif, A. Badhib, H. Ammar, S. Alshehri, M. Kalkatawi, and A. Imine, "Credit card fraud detection in the era of disruptive technologies: A systematic review," *J. King Saud Univ. Inf. Sci.*, vol. 35, no. 1, pp. 145–174, 2023.
3. S. Makki, Z. Assaghir, Y. Taher, R. Haque, M.-S. Hacid, and H. Zeineddine, "An experimental study with imbalanced classification approaches for credit card fraud detection," *IEEE Access*, vol. 7, pp. 93010–93022, 2019.
4. E. Ileberi, Y. Sun, and Z. Wang, "A machine learning based credit card fraud detection using the GA algorithm for feature selection," *J. Big Data*, vol. 9, no. 1, p. 24, 2022.
5. H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, 2009.
6. N. V Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.
7. P. Hajek, M. Z. Abedin, and U. Sivarajah, "Fraud detection in mobile payment systems using an XGBoost-based framework," *Inf. Syst. Front.*, vol. 25, no. 5, pp. 1985–2003, 2023.
8. I. D. Mienye and Y. Sun, "A deep learning ensemble with data resampling for credit card fraud detection," *IEEE Access*, vol. 11, pp. 30628–30638, 2023.
9. E. Strelcenia and S. Prakoonwit, "Improving classification performance in credit card fraud detection by using new data augmentation," *AI*, vol. 4, no. 1, pp. 172–198, 2023.
10. M. Piccininni, M. Wechsung, B. Van Calster, J. L. Rohmann, S. Konigorski, and M. van Smeden, "Understanding random resampling techniques for class imbalance correction and their consequences on calibration and discrimination of clinical risk prediction models," *J. Biomed. Inform.*, vol. 155, p. 104666, 2024.
11. A. Mishra, R. L. McClelland, L. Y. T. Inoue, and K. F. Kerr, "Recalibration methods for improved clinical utility of risk scores," *Med. Decis. Mak.*, vol. 42, no. 4, pp. 500–512, 2022.

12. M. Kull, T. M. Silva Filho, and P. Flach, "Beyond sigmoids: How to obtain well-calibrated probabilities from binary classifiers with beta calibration," *Electron. J. Stat.*, vol. 11, no. 2, pp. 5052–5080, 2017.
13. C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *International conference on machine learning*, PMLR, 2017, pp. 1321–1330.
14. G. Menardi and N. Torelli, "Training and assessing classification rules with imbalanced data," *Data Min. Knowl. Discov.*, vol. 28, no. 1, pp. 92–122, 2014.
15. A. Dal Pozzolo, O. Caelen, R. A. Johnson, and G. Bontempi, "Calibrating probability with undersampling for unbalanced classification," in *Proceedings of the IEEE Symposium Series on Computational Intelligence (SSCI)*, 2015, pp. 159–166.
16. J. L. Leevy, J. M. Johnson, J. Hancock, and T. M. Khoshgoftaar, "Threshold optimization and random undersampling for imbalanced credit card data," *J. Big Data*, vol. 10, no. 1, p. 58, 2023, doi: 10.1186/s40537-023-00738-z.
17. T. Silva Filho, H. Song, M. Perello-Nieto, R. Santos-Rodriguez, M. Kull, and P. Flach, "Classifier calibration: a survey on how to assess and improve predicted class probabilities," *Mach. Learn.*, vol. 112, no. 9, pp. 3211–3260, 2023, doi: 10.1007/s10994-023-06336-7.
18. G. Douzas, F. Bacao, and F. Last, "Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE," *Inf. Sci. (Nijmegen)*, vol. 465, pp. 1–20, 2018.
19. C. Elkan, "The foundations of cost-sensitive learning," in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2001, pp. 973–978.
20. A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," in *Proceedings of the 22nd international conference on Machine learning*, 2005, pp. 625–632. doi: 10.1145/1102351.1102430.
21. M. P. Naeini, G. Cooper, and M. Hauskrecht, "Obtaining well calibrated probabilities using bayesian binning," in *Proceedings of the AAAI conference on artificial intelligence*, 2015.
22. G. W. Brier, "Verification of forecasts expressed in terms of probability," *Mon. Weather Rev.*, vol. 78, no. 1, pp. 1–3, 1950.
23. J. Nixon, M. W. Dusenberry, L. Zhang, G. Jerfel, and D. Tran, "Measuring calibration in deep learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019, pp. 38–41.
24. J. C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in Large Margin Classifiers*, MIT Press, 2000, pp. 61–74.
25. B. Zadrozny and C. Elkan, "Transforming classifier scores into accurate multiclass probability estimates," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2002, pp. 694–699.
26. S. Rajaraman, P. Ganesan, and S. Antani, "Deep learning model calibration for improving performance in class-imbalanced medical image classification tasks," *PLoS One*, vol. 17, no. 1, p. e0262838, 2022.
27. T. Guilbert, O. Caelen, A. Chirita, and M. Saeens, "Calibration methods in imbalanced binary classification," *Ann. Math. Artif. Intell.*, vol. 92, pp. 1319–1352, 2024.
28. Z. Liu, "The hidden cost of resampling: How imbalance correction degrades probability calibration in tree ensembles," 2026. doi: 10.48550/arXiv.2606.29720.
29. F. Lin, Y. Dai, and Y. Liu, "On probability estimation for unbalanced classification," *J. Stat. Theory Pract.*, vol. 20, p. 57, 2026, doi: 10.1007/s42519-026-00572-5.
30. kartik2112, "Credit Card Transactions Fraud Detection Dataset." Kaggle, Aug. 13, 2020. [Online]. Available: <https://www.kaggle.com/datasets/kartik2112/fraud-detection>