e-ISSN 2549-7472
Edumatic: Jurnal Pendidikan Informatika
Volume 10, Issue 1, **21-29.**
https://e-journal.hamzanwadi.ac.id/index.php/edumatic
DOI: 10.29408/edumatic.v10i1.33272

Revisiting Resampling Strategies under Extreme Class Imbalance: Evidence from Large-Scale Online Payment Fraud Detection

Mursyid Ardiansyah [1,*], Ali Asgar Zainal Abidin [2]

[1] Universiti Malaysia Perlis, Malaysia
[2] Institut Teknologi Sains dan Bisnis Muhammadiyah Selayar, Indonesia
[*] Corresponding author: Mursyid Ardiansyah, Universiti Malaysia Perlis, Malaysia
✉ mursyid@studentmail.unimap.edu.my

Received: 2 December 2025 | Revised: 15 January | Accepted: 22 February 2026 | Published: 3 March 2026

Abstract
Extreme class imbalance in online payment fraud detection creates an accuracy paradox and an operational risk in which improving fraud capture can generate costly false alarms. This study uses a quantitative, experiment-based design to evaluate the operational impact of common resampling strategies under extreme skew using interpretable linear decision rules. The Online Payments Fraud dataset (6.36 million transactions) from Kaggle is analysed using six monetary balance/amount variables (amount, oldbalanceOrg, newbalanceOrig, oldbalanceDest, newbalanceDest) plus the rule-based isFlaggedFraud indicator to predict the isFraud label. Five training variants (no resampling, ROS, RUS, SMOTE, ADASYN) are compared with two linear decision rules: an ordinary least squares linear scoring model (thresholded at 0.5) and a linear SVM, using a leakage-free protocol in which resampling is applied only to the 80% training split and performance is assessed on an untouched, highly imbalanced 20% test set. The findings indicate that LinReg–RUS achieves the most balanced operating point (Precision 65.938%, Recall 47.718%, F1 55.367%, ROC-AUC 98.720%), whereas ADASYN increases recall but collapses precision (~2.1%), yielding F1 ≈4.17%. These results contribute controlled, large-scale evidence that under extreme imbalance, simpler resampling–model combinations can provide more deployable precision–recall trade-offs than aggressive synthetic sampling, supporting interpretable baselines for capacity-constrained payment screening.

Keywords: adasyn; fraud detection; random oversampling; random undersampling; smote

To cite this article: Ardiansyah, M., & Abidin, A. A. Z. (2026). Revisiting Resampling Strategies under Extreme Class Imbalance: Evidence from Large-Scale Online Payment Fraud Detection. *Edumatic: Jurnal Pendidikan Informatika*, *10*(1), 21–29. https://doi.org/10.29408/edumatic.v10i1.33272

INTRODUCTION

Online payment fraud screening operates under two coupled constraints: an extremely low fraud base rate and a limited operational capacity to investigate alerts. Under these conditions, increasing recall often inflates false positives, so performance must be interpreted through workload-sensitive metrics rather than headline accuracy Cherif et al. (2023), Hernandez Aros et al. (2024), and Vanini et al. (2023). For real payment operations, the scientific tension is not whether fraud can be detected at all, but whether minority capture can be improved without making the alert stream unmanageable, placing precision–recall trade-offs at the centre of evaluation rather than headline accuracy Alamri & Ykhlef (2022). From the outset, this study focuses on linear, interpretable decision rules not as a limitation, but as a

deliberate design choice aligned with deployment realities: auditability, regulatory defensibility, stable decision boundaries, and cost-aware alerting in high-volume screening environments (Höppner et al., 2022).

It is well established that fraud detection pipelines must simultaneously address scale (millions of records and strict latency budgets) and skew (rare positives), motivating widespread use of imbalance-aware strategies such as resampling, threshold adjustment, and hybrid modelling (Azim Mim et al., 2024; Bounab et al., 2024). Recent reviews also show an accelerating shift toward deep and ensemble architectures, frequently justified by predictive gains but often accompanied by reduced transparency and weaker interpretability for operational auditing Baisholan et al. (2025). At the same time, applied evidence across financial and related high-risk domains indicates that simpler models can remain competitive when evaluation is aligned with deployment costs, error asymmetry, and investigation capacity constraints Höppner et al. (2022).

What remains debated is the reliability of common imbalance remedies across dataset regimes and model classes. Random oversampling and random undersampling are often treated as "baseline preprocessing," yet both can materially change decision boundaries either by duplicating minority patterns or discarding majority structure Brishti et al. (2025), and Carvalho et al. (2025). Synthetic methods such as SMOTE and ADASYN are widely reported to improve minority sensitivity, but their effects depend strongly on feature geometry, class overlap, and classifier interaction; in some settings, synthetic points can amplify noise or distort local neighbourhood structure, pushing models toward unstable alerting behaviour Du et al. (2023), Khalid et al. (2024), and Zhao & Bai (2022). Controlled comparisons in non-financial domains likewise show that SMOTE/ADASYN can help some learners while producing limited or inconsistent gains for others, reinforcing that resampling is not a neutral step but an intervention whose impact depends on method-model coupling Hidayat et al. (2021).

A further debate concerns how performance should be interpreted under extreme skew. ROC-AUC is often reported as a primary indicator, yet methodological work highlights that AUC-style summaries can remain high even when operational precision is unacceptable, especially when the base rate is tiny; evaluation therefore needs to foreground confusion-matrix structure and precision-recall behaviour to reflect screening workload and customer impact Richardson et al. (2024), and Van Calster et al. (2025).

Most fraud studies evaluate resampling together with complex models, frequently report results on balanced or partially rebalanced test sets, and emphasise ROC-AUC rather than the operational burden created by precision collapse. In contrast, comparatively few studies isolate the interaction between resampling strategies and model behaviour using interpretable linear decision rules, evaluate on an untouched and extremely imbalanced test set, and explicitly quantify how precision collapse translates into operational risk (excessive false alarms). As a result, it remains unresolved whether model complexity is necessary to achieve a usable precision-recall balance under extreme imbalance, or whether simpler resampling-model combinations can provide more stable and deployable screening performance at scale.

This study systematically evaluates the operational impact of four widely used resampling strategies (ROS, RUS, SMOTE, and ADASYN) against an unresampled baseline under extreme class imbalance. To isolate resampling model coupling (rather than architectural complexity), the experiments compare two interpretable linear decision rules: a linear scoring model with a fixed operating threshold and a linear SVM using margin-based optimisation. All models are evaluated on an untouched, highly imbalanced test set so that the reported precision–recall behaviour directly reflects deployability under realistic fraud prevalence.

METHOD

The study adopted a quantitative, experiment-based design using secondary transactional data. Data were taken from the Online Payments Fraud Detection Dataset on Kaggle (https://www.kaggle.com/datasets/rupakroy/online-payments-fraud-detection-dataset), which contains anonymised online payment records labelled as fraudulent or legitimate. Experiments were conducted offline in Python 3 using scikit-learn and imbalanced-learn with fixed random seeds for reproducibility.

Table 1 clarifies the role of each dataset field in the study by distinguishing retained predictors, excluded variables, and the target label. Predictors are restricted to balance- and amount-dynamics (amount, oldbalanceOrg, newbalanceOrig, oldbalanceDest, newbalanceDest) plus the system flag (isFlaggedFraud) to maintain interpretability and reduce sensitivity to high-dimensional encodings. Identifier fields (nameOrig, nameDest) and transaction type are excluded because they require large categorical representations that may hinder generalisability and obscure the analysis of monetary-behaviour signals in linear decision rules.

Table 1. Feature description.

Variable	Type	Description
step	Integer	Time-step index in the dataset
type	Categorical	Transaction type (excluded to avoid high-dimensional encoding)
amount	Numeric	Transaction amount
nameOrig	String ID	Sender identifier (excluded; non-generalizable ID)
oldbalanceOrg	Numeric	Sender balance before transaction
newbalanceOrig	Numeric	Sender balance after transaction
nameDest	String ID	Receiver identifier (excluded; non-generalizable ID)
oldbalanceDest	Numeric	Receiver balance before transaction
newbalanceDest	Numeric	Receiver balance after transaction
isFraud	Binary	Ground-truth label (1 = fraud, 0 = legitimate)
isFlaggedFraud	Binary	Rule-based system flag (1 = flagged)

Predictors were limited to amount, oldbalanceOrg, newbalanceOrig, oldbalanceDest, newbalanceDest, and isFlaggedFraud, with isFraud as the target label. Records with missing values in retained fields were removed. Numeric predictors were standardised using z-score scaling fit on the training split and applied to both training and test splits. The cleaned dataset was split into training (80%) and test (20%) subsets using stratified sampling to preserve the extreme class imbalance. To prevent information leakage, resampling was applied only to the training subset, while the test subset remained untouched and imbalanced for evaluation Chen et al. (2024).

Five training variants were created from the same base training split: (i) no resampling (baseline), (ii) Random Over-Sampling (ROS), (iii) Random Under-Sampling (RUS), (iv) SMOTE, and (v) ADASYN (imbalanced-learn). SMOTE and ADASYN used default settings (k_neighbors = 5, sampling_strategy = “auto”) with random_state = 42. For each variant, two linear decision rules were trained: (1) an ordinary least squares linear scoring model producing continuous scores thresholded at 0.5, and (2) a linear SVM (LinearSVC) using decision_function scores. Performance on the untouched test set was reported using accuracy, precision, recall, and F1-score with confusion matrices; ROC curves and ROC-AUC were computed from continuous scores. A single stratified 80/20 split was used for computational practicality and is reported as a limitation motivating repeated splits or cross-validation.

RESULTS AND DISCUSSION

Results

The results are interpreted under a highly imbalanced base rate that reflects real-world screening conditions. As reported in Table 2, fraudulent transactions represent only a tiny fraction of the dataset, which makes accuracy dominated by the majority class and therefore insufficient for assessing fraud detection quality. Consequently, performance comparisons emphasise precision, recall, F1-score, confusion-matrix structure, and ROC-AUC, where precision reflects alert burden and recall reflects fraud capture. This framing ensures that improvements are discussed in terms of deployability rather than apparent gains driven by class prevalence.

Table 4 reports performance for two linear decision rules trained with five variants (no resampling, ROS, RUS, SMOTE, and ADASYN). Among all configurations, LinReg-RUS achieves the highest F1-score (55.367%) with Precision 65.938%, Recall 47.718%, and ROC-AUC 98.720%. The SVM baseline yields the highest precision (97.512%) but lower recall (35.788%), resulting in F1-score 52.360% and ROC-AUC 98.552%. In fraud studies with extreme class imbalance, the fraud-class F1 commonly reported for standard baselines can be very low (around ~0.01–0.10 for some baseline models), and even stronger benchmarks may remain near ~0.29–0.34; therefore, an F1 above 0.55 in Table 4 can be interpreted as a meaningful operational improvement in fraud capture while maintaining a still-usable precision level.

Table 2. Class distribution

isFraud	Total	Percentage
0	6,354,407	99.9
1	8,213	0.1

As shown in Table 2, fraud cases constitute only 8,213 of 6,362,620 transactions (0.1%), confirming an extreme skew in which the majority class dominates the data-generating process. Under this prevalence, a classifier can obtain very high accuracy by predicting nearly all cases as legitimate, even while failing to identify a meaningful portion of fraud events. This class distribution therefore motivates the use of imbalance-handling strategies during training, because the learning objective otherwise becomes overly aligned with majority-class fit and may underweight minority patterns. It also requires evaluation metrics that explicitly reflect minority detection and alert burden particularly precision, recall, F1-score, and confusion-matrix structure rather than relying on headline accuracy alone. In practical screening settings, this imbalance implies that even small false-positive rates can translate into large absolute numbers of alerts, making the precision–recall trade-off central to operational usefulness.

Table 3. Split data

Split	Total Data	Percentage (%)
Train	5,090,096	80
Test	1,272,524	20

The 80/20 stratified split in Table 3 provides a practical balance between model fitting capacity and unbiased evaluation at scale. With 5,090,096 records allocated to training, the resampling procedures (ROS, RUS, SMOTE, and ADASYN) can be applied without compromising statistical stability, enabling the models to learn minority-class structure from a sufficiently large sample. At the same time, retaining 1,272,524 records for testing preserves a large, independent hold-out set that better reflects the distribution of unseen transactions in real screening scenarios. Stratification ensures that the extreme class imbalance is preserved across

both partitions, preventing accidental enrichment of fraud cases in either split that could inflate performance estimates. This design supports a leakage-aware protocol in which all preprocessing and resampling decisions are constrained to the training subset, while the test subset remains untouched to measure generalisation under deployment-like conditions.

Table 4. Model performance

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
LinReg-RUS	99.901%	65.938%	47.718%	55.367%	98.720%
LinReg-ROS	99.892%	59.559%	51.004%	54.951%	98.223%
LinReg-SMOTE	99.890%	58.397%	51.430%	54.693%	97.889%
SVM-Baseline	99.916%	97.512%	35.788%	52.360%	98.552%
LinReg-Baseline	99.883%	100.000%	9.677%	17.647%	95.275%
SVM-RUS	97.415%	3.920%	80.889%	7.477%	98.339%
SVM-SMOTE	97.179%	3.813%	86.062%	7.302%	98.719%
SVM-ROS	96.885%	3.404%	84.480%	6.544%	98.488%
LinReg-ADASYN	94.805%	2.139%	87.705%	4.177%	97.849%
SVM-ADASYN	94.301%	2.131%	96.044%	4.170%	98.689%

Table 4 highlights distinct operating behaviors, synthetic oversampling consistently increases recall but destabilizes precision, whereas random resampling yields a more balanced operational profile. Specifically, LinReg with light resampling (RUS/ROS/SMOTE) achieves the best balance, with F1-scores around 54.693-55.367% and substantially higher precision (58.397-65.938%) than any resampled SVM variant. By contrast, resampled SVM variants push recall to 80.889-96.044% but collapse precision to 2.131-3.920%, yielding very low F1-scores (4.170-7.477%), many false alarms. ADASYN shows the same failure mode for both model families, indicating that aggressive synthetic sampling is not suitable for this dataset under the fixed operating point.

Under extreme skew, resampling expands the effective minority region seen during training. For linear decision rules, this can shift the decision boundary toward the majority class (boundary shift) to reduce false negatives. However, when minority and majority feature regions partially overlap, synthetic oversampling can enlarge the minority support in directions that intrude into majority space (minority space expansion and class overlap), increasing ambiguous boundary regions. In a margin-based classifier such as linear SVM, this overlap can distort the margin (margin distortion), causing many majority points near the boundary to be labeled as fraud, which explains the sharp precision collapse observed for SVM-ROS/SMOTE/ADASYN despite high recall. In contrast, the LinReg scoring rule produces a smoother continuous score surface; with RUS/ROS/SMOTE it shifts toward improved minority detection while maintaining a more controlled false-positive rate, yielding the best F1-score. ADASYN is particularly prone to generating samples in difficult or borderline areas, which can introduce synthetic noise in heavily overlapping regions and further inflate false positives for both model families.

Figures 1 and 2 show ROC curves for LinReg-RUS and the SVM baseline on the untouched imbalanced test set. Both curves indicate strong rank separability (AUC = 0.98720 for LinReg-RUS and AUC = 0.98552 for the SVM baseline). However, under extreme skew, high ROC-AUC can coexist with poor precision at a chosen threshold; thus, ROC-AUC is interpreted jointly with precision, recall, F1-score, and confusion-matrix structure.

Figures 3 and 4 compare confusion matrices of the SVM baseline and LinReg-RUS. The SVM baseline classifies 1,270,866 legitimate transactions correctly and produces only 15 false positives, but it detects 588 fraud cases and misses 1,055 (higher precision, lower recall). LinReg-RUS detects more fraud (784 true positives) and misses fewer (859 false negatives)

but increases false positives to 405 (a more aggressive boundary). This illustrates why LinReg-RUS achieves the best F1-score: it increases fraud capture while keeping false alarms within a more controlled range than resampled SVM variants.

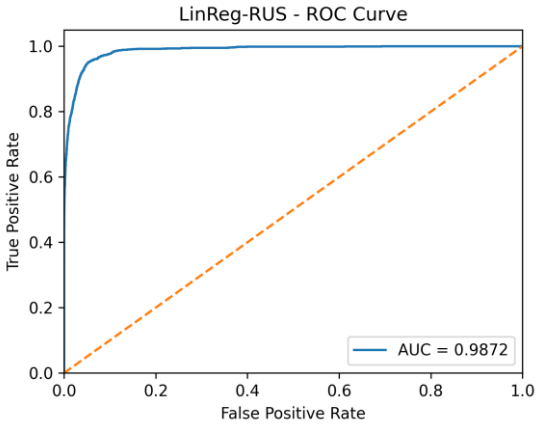

Figure 1. ROC curve of LinReg-RUS

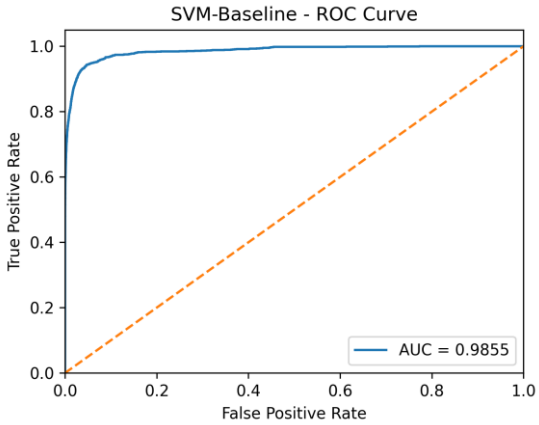

Figure 2. ROC curve of SVM baseline

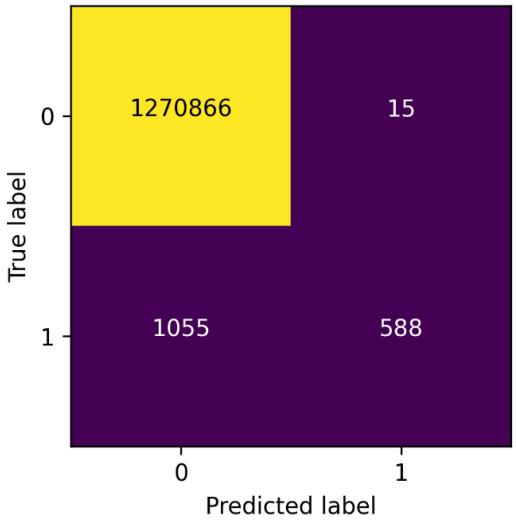

Figure 3. Confusion matrix of SVM baseline (no resampling).

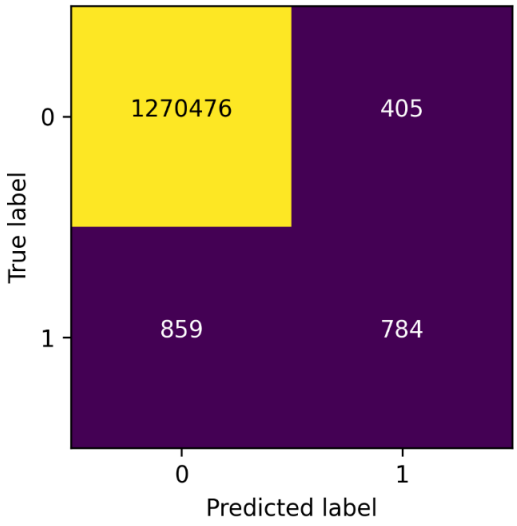

Figure 4. Confusion matrix of LinReg-RUS.

Discussion

This study shows that, under extreme class imbalance in large-scale online payment transactions, simple random resampling paired with an interpretable linear scoring rule can provide a more balanced operational profile than aggressive synthetic oversampling paired with a margin-based classifier. In practical screening terms, moderate resampling supports improved fraud capture without generating an unmanageable volume of false alarms, whereas aggressive synthetic sampling increases sensitivity but can destabilise precision and thereby degrade alert usability.

A central mechanism in imbalanced learning is that resampling can move the decision boundary by changing which regions of the feature space are densely represented during training. The findings indicate a controlled generalisation: when class separability is limited, increasing minority density via synthetic generation may unintentionally reshape local class topology. In this view, resampling alters the empirical risk landscape; where minority and majority regions partially overlap, synthetic expansion can intensify overlap in boundary-adjacent neighbourhoods and shift the effective decision region toward the majority class.

Consequently, improvements in ranking-based separability (ROC-AUC) do not necessarily yield an operationally acceptable decision rule, because the selected operating threshold interacts with boundary geometry and the underlying fraud base rate.

The findings align with the broader observation that oversampling gains are not universal and depend on the learner–data interaction, with resampling effects varying across feature structures and classifiers Taskiran et al. (2025). They also refine comparative accounts of SMOTE versus ADASYN by suggesting that ADASYN's adaptive synthesis can amplify borderline complexity in heavily skewed settings, increasing local overlap and making precision collapse more likely when the minority region is thin and entangled with majority neighbourhoods Wah et al. (2023). At the same time, the outcomes complicate the narrative implied in complex fraud-detection frameworks such as stacking ensembles with extensive feature engineering by indicating that operational precision may worsen unless alert capacity and boundary effects are explicitly constrained, even when overall sensitivity improves (Wu et al., 2023; Zeng et al., 2025).

A notable and potentially counter-intuitive pattern is that configurations producing high recall can still yield poor operational performance due to severe precision collapse. Mechanistically, this can be explained by minority manifold expansion into boundary-adjacent majority regions and decision boundary distortion, especially under synthetic sampling that generates points in ambiguous neighbourhoods. In a margin-based classifier, these effects can manifest margin instability, where the separating hyperplane is pulled by synthetic or borderline points, causing many majority instances to be labelled as fraud. This provides a coherent explanation for why synthetic oversampling can appear beneficial under recall-centric evaluation while becoming operationally unattractive under precision-recall trade-offs.

The study contributes a large-scale, controlled evaluation that isolates resampling-model coupling under extreme imbalance using interpretable linear decision rules and an untouched imbalanced test set. Beyond reporting performance differences, the discussion clarifies a conceptual point: resampling is not merely a preprocessing convenience but a geometry-altering intervention whose operational consequences depend on classifier inductive bias (scoring versus margin optimisation) and the overlap structure of the feature space. By challenging the assumption that synthetic oversampling and more complex classifiers are consistently superior, the study strengthens empirical evidence for "deployable simplicity" as a defensible baseline under extreme skew.

In operational payment screening, model selection should be guided by a joint objective: maximising fraud capture while maintaining an alert volume that remains within feasible investigation capacity. Interpretable linear scoring rules paired with moderate resampling can strengthen governance because the decision logic and thresholds are easier to audit, communicate, and calibrate important requirements in regulated environments that demand explainable decisions and stable threshold policies. By contrast, configurations that prioritise recall at the expense of precision can generate a disproportionate false-alarm burden, increasing investigation costs and customer friction, and are therefore unsuitable unless downstream triage capacity or explicit cost-sensitive constraints are incorporated.

Several limitations should be noted. First, the evaluation uses a single stratified split for computational practicality; repeated splits, temporal validation, or cross-validation would provide stronger uncertainty estimates and robustness checks. Second, the feature set is intentionally minimal to preserve interpretability; incorporating richer behavioural, temporal, or network-derived features may change overlap geometry and resampling effects. Third, a fixed thresholding approach limits operating-point optimisation; future work should incorporate cost-sensitive threshold selection, calibration, and alert-budget constraints to align evaluation with deployment objectives. Finally, additional analyses of stability under

distribution shift (concept drift) would clarify how resampling-model coupling behaves over time in payment streams.

CONCLUSION

This study examined how common resampling strategies shape operational fraud screening performance under extreme class imbalance on a large-scale online payment dataset using interpretable linear decision rules. The evidence indicates that lightweight random resampling particularly RUS can improve minority detection while keeping false alarms within a more manageable range than aggressive synthetic sampling, which may increase sensitivity at the cost of severe precision degradation. The findings contribute a controlled, large-scale demonstration that resampling-model coupling can dominate perceived performance gains, and they refine the prevailing view that synthetic oversampling paired with stronger classifiers is universally advantageous under extreme skew. At a field level, this implies that high-volume payment monitoring can benefit from transparent scoring rules that support auditing, stable thresholding, and capacity-aware alerting rather than relying solely on algorithmic complexity. Future work should extend this analysis to richer feature representations, repeated splits or time-aware validation, and explicitly cost-sensitive operating-point selection (cost curves or constrained false-positive budgets), alongside calibration and drift robustness to reflect deployment conditions. Ultimately, the results underscore that in high-risk financial systems, deployable simplicity can remain a competitive alternative to algorithmic sophistication.

REFERENCES

Alamri, M., & Ykhlef, M. (2022). Survey of Credit Card Anomaly and Fraud Detection Using Sampling Techniques. *Electronics (Switzerland)*, *11*(23). https://doi.org/10.3390/electronics11234003

Azim Mim, M., Majadi, N., & Mazumder, P. (2024). A soft voting ensemble learning approach for credit card fraud detection. *Heliyon*, *10*(3), e25466. https://doi.org/10.1016/j.heliyon.2024.e25466

Baisholan, N., Dietz, J. E., Gnatyuk, S., Turdalyuly, M., Matson, E. T., & Baisholanova, K. (2025). A Systematic Review of Machine Learning in Credit Card Fraud Detection Under Original Class Imbalance. *Computers*, *14*(10). https://doi.org/10.3390/computers14100437

Bounab, R., Zarour, K., Guelib, B., & Khlifa, N. (2024). Enhancing Medicare Fraud Detection Through Machine Learning: Addressing Class Imbalance With SMOTE-ENN. *IEEE Access*, *12*(3), 54382–54396. https://doi.org/10.1109/ACCESS.2024.3385781

Brishti, F., Zhang, F., Mohammed, S., Bai, L., Wu, F., & Chen, B. (2025). Imbalanced classification with label noise: A systematic review and comparative analysis. *ICT Express*, *11*(6), 1127–1145. https://doi.org/10.1016/j.icte.2025.09.011

Carvalho, M., Pinho, A. J., & Brás, S. (2025). Resampling approaches to handle class imbalance: a review from a data perspective. *Journal of Big Data*, *12*(1). https://doi.org/10.1186/s40537-025-01119-4

Chen, W., Yang, K., Yu, Z., Shi, Y., & Chen, C. L. P. (2024). A survey on imbalanced learning: latest research, applications and future directions. *Artificial Intelligence Review*, *57*(6). https://doi.org/10.1007/s10462-024-10759-6

Cherif, A., Badhib, A., Ammar, H., Alshehri, S., Kalkatawi, M., & Imine, A. (2023). Credit card fraud detection in the era of disruptive technologies: A systematic review. *Journal of King Saud University - Computer and Information Sciences*, *35*(1), 145–174. https://doi.org/10.1016/j.jksuci.2022.11.008

Du, H., Lv, L., Guo, A., & Wang, H. (2023). AutoEncoder and LightGBM for Credit Card Fraud Detection Problems. *Symmetry*, *15*(4). https://doi.org/10.3390/sym15040870

Hernandez Aros, L., Bustamante Molano, L. X., Gutierrez-Portela, F., Moreno Hernandez, J. J., & Rodríguez Barrero, M. S. (2024). Financial fraud detection through the application of machine learning techniques: a literature review. *Humanities and Social Sciences Communications*, *11*(1). https://doi.org/10.1057/s41599-024-03606-0

Hidayat, W., Ardiansyah, M., & Setyanto, A. (2021). Pengaruh Algoritma ADASYN dan SMOTE terhadap Performa Support Vector Machine pada Ketidakseimbangan Dataset Airbnb. *Edumatic: Jurnal Pendidikan Informatika*, *5*(1), 11–20. https://doi.org/10.29408/edumatic.v5i1.3125

Höppner, S., Baesens, B., Verbeke, W., & Verdonck, T. (2022). Instance-dependent cost-sensitive learning for detecting transfer fraud. *European Journal of Operational Research*, *297*(1), 291–300. https://doi.org/10.1016/j.ejor.2021.05.028

Khalid, A. R., Owoh, N., Uthmani, O., Ashawa, M., Osamor, J., & Adejoh, J. (2024). Enhancing Credit Card Fraud Detection: An Ensemble Machine Learning Approach. *Big Data and Cognitive Computing*, *8*(1). https://doi.org/10.3390/bdcc8010006

Richardson, E., Trevizani, R., Greenbaum, J. A., Carter, H., Nielsen, M., & Peters, B. (2024). The receiver operating characteristic curve accurately assesses imbalanced datasets. *Patterns*, *5*(6), 100994. https://doi.org/10.1016/j.patter.2024.100994

Taskiran, S. F., Turkoglu, B., Kaya, E., & Asuroglu, T. (2025). A comprehensive evaluation of oversampling techniques for enhancing text classification performance. *Scientific Reports*, *15*(1). https://doi.org/10.1038/s41598-025-05791-7

Van Calster, B., Collins, G. S., Vickers, A. J., Wynants, L., Kerr, K. F., Barreñada, L., Varoquaux, G., Singh, K., Moons, K. G., Hernandez-Boussard, T., Timmerman, D., McLernon, D. J., van Smeden, M., & Steyerberg, E. W. (2025). Evaluation of performance measures in predictive artificial intelligence models to support medical decisions: overview and guidance. *The Lancet Digital Health*, *7*(12), 100916. https://doi.org/10.1016/j.landig.2025.100916

Vanini, P., Rossi, S., Zvizdic, E., & Domenig, T. (2023). Online payment fraud: from anomaly detection to risk management. *Financial Innovation*, *9*(1). https://doi.org/10.1186/s40854-023-00470-w

Wah, Y. B., Ismail, A., Azid, N. N. N., Jaafar, J., Aziz, I. A., Hasan, M. H., & Zain, J. M. (2023). Machine Learning and Synthetic Minority Oversampling Techniques for Imbalanced Data: Improving Machine Failure Prediction. *Computers, Materials and Continua*, *75*(3), 4821–4841. https://doi.org/10.32604/cmc.2023.034470

Wu, B., Lv, X., Alghamdi, A., Abosaq, H., & Alrizq, M. (2023). Advancement of management information system for discovering fraud in master card based intelligent supervised machine learning and deep learning during SARS-CoV2. *Information Processing & Management*, *60*(2), 103231. https://doi.org/10.1016/j.ipm.2022.103231

Zeng, Q., Lin, L., Jiang, R., Huang, W., & Lin, D. (2025). NNEnsLeG: A novel approach for e-commerce payment fraud detection using ensemble learning and neural networks. *Information Processing & Management*, *62*(1), 103916. https://doi.org/10.1016/j.ipm.2024.103916

Zhao, Z., & Bai, T. (2022). Financial Fraud Detection and Prediction in Listed Companies Using SMOTE and Machine Learning Algorithms. *Entropy*, *24*(8). https://doi.org/10.3390/e24081157