https://doi.org/10.70882/qg6d3k52

SCIENCE FORUM (JOURNAL OF PURE AND APPLIED SCIENCES) 26, (4) (2026), 1415 - 1425

Science Forum (Journal of Pure and Applied Sciences)

Journal Homepage: https://atbuscienceforum.com.ng/index.php/jpas

Big Data-Driven Deep Learning Models for Real-Time Fraud Detection

Nebras Jalel Ibrahim[1], Ahmed K. Abbas[2]

[1]Computer Center, University of Diyala,32001,Diyala, Iraq

[2]Collage of Education for pure science, University of Diyala,32001Diyala,Iraq

Abstract

Fraud detection is a rare-event decision problem: a model can report a high overall accuracy while failing to identify the transactions that matter. We therefore evaluated four classifiers under a chronological protocol using the public Credit Card Fraud Detection dataset released by the Machine Learning Group of Université Libre de Bruxelles and Worldline. The dataset contains 284,807 transactions, including 492 labeled fraud cases (0.1727%). The original order was retained and divided into training (70%), validation (15%), and final test (15%) periods. The comparison includes class-weighted logistic regression, a balanced random forest, cost-weighted HistGradientBoosting, and a three-hidden-layer multilayer perceptron (MLP). Each threshold was chosen on the validation period by maximizing F1 and then frozen before testing. On the final period, the MLP produced the highest F1-score (0.817) and MCC (0.823), with 38 true positives, 3 false positives, and 14 false negatives. Random Forest produced the best average precision (PR-AUC = 0.765) and the highest recall (0.750). Thus, the neural model delivered the strongest selected operating point, while the forest ranked transactions slightly better under PR-AUC. The evidence is useful as a controlled benchmark, not as a claim of production readiness.

Keywords

Credit Card Fraud Detection;
Deep Learning;
Imbalanced Classification;
Temporal Validation;
Big Data;
Real-Time Analytics;
Reproducible Machine Learning.

EISSN: 3156-5190 (Online)
ISSN: 1119-4618 (Print)
JPAS
Faculty of Science, ATBU Bauchi

Corresponding Author: Nebras Jalel Ibrahim, computer Center, University of Diyala,32001, Diyala, Iraq. Email: nebras.jalel@uodiyala.edu.iq

© 2026 Faculty of Science, ATBU Bauchi. All rights reserved

1. Introduction

A payment-fraud model is used in an environment where the positive class is scarce, labels may arrive late, and customer behavior changes over time. The useful output is therefore not a generic accuracy figure; it is a ranking or decision rule that captures fraud without overwhelming analysts or legitimate customers with false alerts. Earlier comparative studies provide supervised baselines for credit-card fraud [1], [2], and recent reviews describe the expanding use of neural networks, sequence models, imbalance strategies, and specialized evaluation procedures [3].

The ULB–Worldline dataset offers a large labeled benchmark with a documented fraud rate of 0.1727% [31]. Its 28 principal-component variables are anonymized; only Time and Amount retain an immediately recognizable interpretation. The anonymization is valuable for privacy, but it also means that the benchmark cannot answer questions about merchant category, device linkage, customer history, or other features normally available in a production system.

The study asks three practical questions: how the MLP compares with strong tabular baselines when the chronology is respected; which model offers the most useful balance between ranking quality and thresholded capture; and whether the comparison can be reproduced from the supplied data and code. The contribution is consequently methodological rather than promotional: a temporal holdout, validation-only threshold selection, a metric set suited to extreme imbalance, explicit confusion counts and runtime measurements, and an executable record of the experiment.

2. Related Work

The relevant IEEE literature falls into four overlapping strands. The first consists of comparative and application studies of logistic regression, tree ensembles, neural networks, and related supervised learners for credit-card and online-payment fraud [1], [2], [10]–[13], [17]–[20]. Together, these studies support a comparative design rather than an assumption that one model family is always preferable.

The second strand addresses the dominant statistical difficulty: fraud is a very small minority class. Studies have compared oversampling and undersampling directly [5] and have proposed hybrid sampling schemes [6], [7]. Other work examines neural networks [8], Random Forest [9], two-stage thresholding [14], and noise-aware undersampling [30]. This literature is the reason accuracy is not used as the principal decision metric here.

Third, deep learning and explainability have received increasing attention. A recent IEEE Access review summarizes CNN, RNN, LSTM, GRU, imbalance strategies, datasets, and evaluation challenges [3]. Online-payment systems have combined deep feature extraction, recurrent components, autoencoders, and resampling [4]. Explainable machine-learning approaches have been applied to credit-card fraud to make high-impact decisions more auditable [16].

Fourth, graph-based and real-time approaches model the relational structure among accounts, cards, devices, merchants, and transactions. GNN models have been proposed for financial fraud [21], [22], credit-card fraud [29], and ensembles of GNN architectures [23]. Adaptive graph learning with reinforcement learning has also been investigated [24]. Other IEEE studies explicitly address anomaly detection and risk assessment [25], comprehensive fraud-prevention implementations [26], and real-time fraud detection in online transactions [27], [28]. The present study is deliberately narrower: it evaluates a deep MLP on a public tabular benchmark and does not claim to solve graph construction, online feature serving, or concept drift.

3. Dataset and Experimental Design

3.1 Dataset

The data were downloaded from the official Kaggle record associated with the MLG-ULB publisher [31]. The extracted file contains 284,807 rows and 31 columns: Time, Amount, V1–V28, and Class. There are 284,315 legitimate transactions and 492 fraud transactions. The fraud prevalence is 0.1727%. Time ranges from 0 to 172,792 seconds, and the supplied file was verified to be monotonically increasing in Time. Figure 1 shows the class distribution.

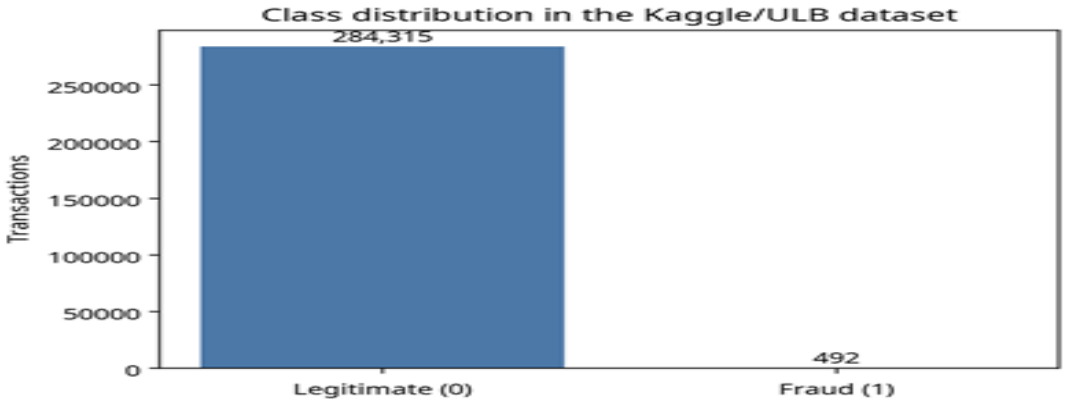


Figure 1. Legitimate and fraudulent transactions in the full dataset.

3.2 Temporal split and leakage control.

The file order was preserved without shuffling. The first 70% (199,364 rows; 384 frauds) formed the training set, the next 15% (42,721 rows; 56 frauds) formed the validation set, and the final 15% (42,722 rows; 52 frauds) formed the test set. Feature scaling was fitted on training data only inside a pipeline. Model and threshold decisions were made without inspecting test labels. This design reflects the concern that random splits can create temporal leakage and over-optimistic estimates in fraud data [30].

3.3 Models. Logistic Regression used StandardScaler and balanced class weights. Random Forest used 200 trees, square-root feature subsampling, minimum leaf size two, and balanced-subset weights. HistGradientBoosting used 250 boosting iterations, learning rate 0.08, 31 leaf nodes, L2 regularization of 1.0, and balanced sample weights. The Deep MLP used hidden layers of 128, 64, and 32 ReLU units, Adam optimization, batch size 1024, learning rate 0.001, early stopping, validation fraction 0.10, and a maximum of 40 epochs. All stochastic models used seed 42. The complete experimental and decision pipeline is illustrated in Figure 2.

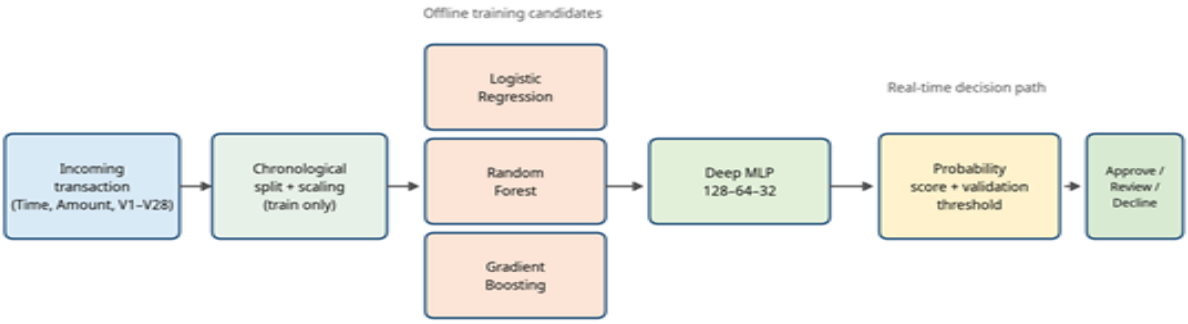


Figure 2. Data flow from the incoming transaction through preprocessing, candidate models, probability scoring, and the final operating decision.

3.4 Threshold and metrics. Each model produced a fraud probability.

The threshold maximizing validation F1 was then frozen and applied to the test period. We report average precision (PR-AUC), ROC-AUC, precision, recall, F1, balanced accuracy, MCC, TP, FP, FN, TN, training time, and approximate per-row inference time. PR-AUC is emphasized because the positive class is extremely rare and the dataset description itself recommends precision–recall evaluation [31].

4. Results

The test-period comparison is summarized in Table 1 and the confusion counts are shown in Table II and Fig. 6. The MLP produced the strongest thresholded result: F1 = 0.817 and MCC = 0.823. At its validation-selected threshold of 0.333, it detected 38 of the 52 fraud cases, produced 3 false positives, and missed 14 cases. Random Forest captured the most fraud cases (39) and achieved the best PR-AUC (0.765), but it also generated 10 false positives. Logistic Regression was the most precise model (0.943) and detected 33 fraud cases. HistGradientBoosting occupied an intermediate position in the operating-point comparison and recorded the lowest PR-AUC (0.724).

Table 1. Test-period performance and operating-point counts.

Model	PR-AUC	ROC-AUC	Precision	Recall	F1	MCC	TP	FP	FN
Logistic Regression	0.736	0.977	0.943	0.635	0.759	0.773	33	2	19
Random Forest	0.765	0.937	0.796	0.750	0.772	0.772	39	10	13
HistGradientBoosting	0.724	0.948	0.818	0.692	0.750	0.752	36	8	16
Deep MLP	0.759	0.958	0.927	0.731	0.817	0.823	38	3	14

The comparison does not produce a single winner under every criterion. The MLP is strongest when the validation-selected operating point is judged by F1 and MCC; Random Forest is stronger as a ranking

model under average precision and recall. The distinction is operationally meaningful. A review queue may favor recall and PR-AUC, whereas an automated step-up rule may place a higher value on precision and a small false-positive count.The test period contains only 52 fraud cases. Consequently, a change of only a few decisions can materially change recall, F1, or MCC.

Figure 3. Precision–recall curves on the temporal test set (52 fraud cases). Average precision (AP) is shown in the legend. Curves converge near recall ≈ 0.75, beyond which precision falls sharply for all models. Step-like behavior reflects the small number of positive cases; one fraud corresponds to roughly 1.9 percentage points of recall.

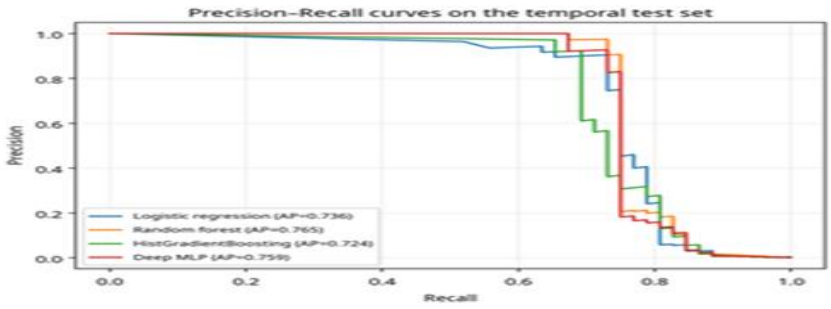


Figure 3. Precision–recall curves on the temporal test set; labels show average precision.

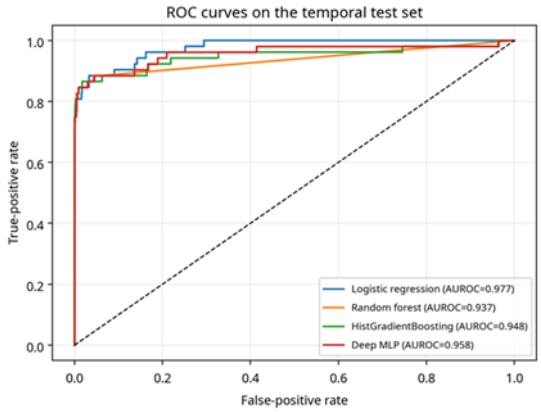


Figure 4. ROC curves on the temporal test set.

Figure 4. ROC curves on the temporal test set (52 fraud cases). AUROC is shown in the legend; the dashed line indicates chance. Curves overlap closely at low false-positive rates, where operating points for fraud screening lie, and differ mainly at high false-positive rates. AUROC can be optimistic under extreme class imbalance; precision-recall results (Figure 5) are therefore emphasized.

Figure 5. Key test-set metrics (PR-AUC, F1, and MCC) for the four models on the temporal test period (52 fraud cases). The Deep MLP attains the highest F1 and MCC, whereas Random Forest attains the highest PR-AUC. Differences among models are small relative to sampling uncertainty; error bars show 95% bootstrap intervals.

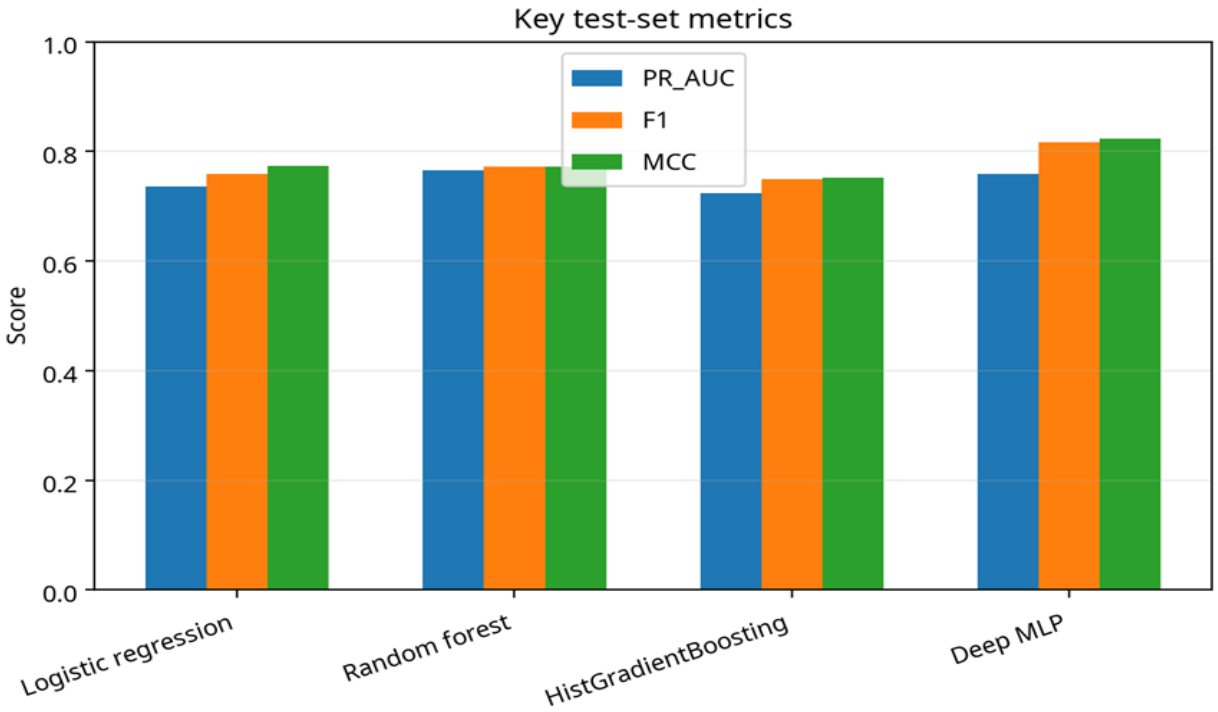


Figure 5. Test PR-AUC, F1, and MCC across the four models.

Table 2. Confusion-matrix counts on the chronological test period

Model	TN	FP	FN	TP
Logistic Regression	42668	2	19	33
Random Forest	42660	10	13	39
HistGradientBoosting	42662	8	16	36
Deep MLP	42667	3	14	38

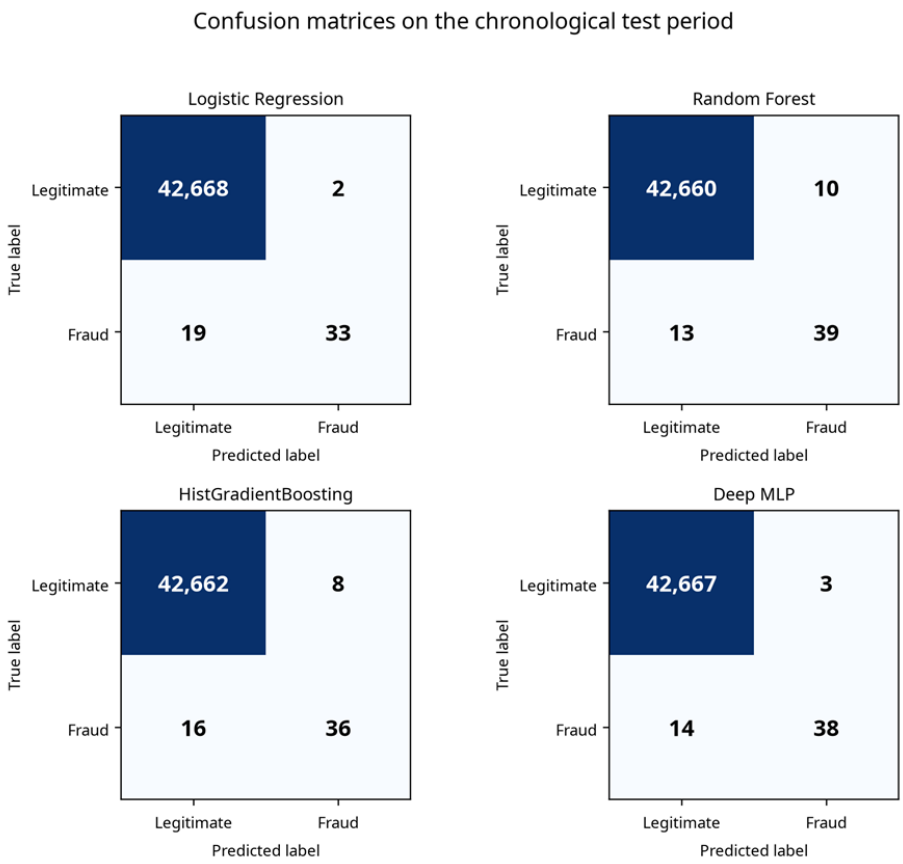


Figure 6. Confusion matrices on the chronological test period; each panel reports TN, FP, FN, and TP counts.

Figure 6. Confusion matrices on the chronological test period (42,722 transactions, 52 frauds). Decision thresholds were selected on the validation period by maximizing F1. The Deep MLP yields 38 true positives, 3 false positives, and 14 false negatives; Random Forest detects one additional fraud (39) at the cost of seven more false alarms (10).

5. Discussion

5.1 Research implications.

The chronological split deliberately treats later transactions as future observations. This reduces the risk of mixing information from adjacent periods, although it also leaves only 52 fraud cases for final testing and therefore increases uncertainty. For a real-time application, that trade-off is preferable to reporting a randomly shuffled estimate that may be easier to obtain but harder to interpret operationally.

5.2 Operational implications.

The measured CPU training times were 0.88 s for Logistic Regression, 23.91 s for Random Forest, 9.88 s for HistGradientBoosting, and 4.09 s for the Deep MLP in the execution environment. Approximate test inference times per row were 0.00036 ms, 0.00332 ms, 0.00146 ms, and 0.00047 ms, respectively. These are local benchmark measurements, not end-to-end service-level measurements: they exclude network, feature-store, serialization, queueing, and monitoring overhead.

5.3 Why accuracy is excluded.

With a fraud prevalence of 0.1727%, an all-legitimate classifier would appear highly accurate while having zero fraud recall. PR-AUC, MCC, and thresholded confusion counts expose the operational trade-off more clearly. This is consistent with the focus on imbalance handling in IEEE studies [5]–[9], [14], [30].

5.4 Real-time limitations.

The offline experiment does not reproduce a complete transaction-scoring service. A deployed system would need online velocity features, account and device relations, delayed-label handling, drift monitoring, calibration, model versioning, audit trails, and a human-review path. Graph methods [21]–[24], [29] may also capture fraud rings that are invisible in this anonymized tabular representation. Accordingly, the real-time studies [27], [28] should be read as addressing an engineering and monitoring problem that cannot be settled by offline classification metrics alone.

6. Limitations and Ethical Considerations

Four limitations should guide the interpretation of the results. First, the observations cover only two days in September 2013. Second, V1–V28 are anonymized PCA components and do not expose their business meaning. Third, the final test period contains only 52 fraud cases, so the operating-point estimates have limited statistical precision. Fourth, the study has no external bank, geography, product, or later time-period validation. The results therefore support a reproducible benchmark comparison, not a claim of cross-bank generalization or production readiness.

Fraud decisions can deny legitimate customers access to payment services. A deployed model should not be the sole basis for irreversible rejection; it should support risk-based authentication or review with an appeal path. Explainability methods [16] and auditable reason codes are especially important when automated scores influence customer outcomes. Any deployment would also require privacy protection, bias testing, access control, secure logging, and compliance review.

8. Conclusion

This study compared a deep MLP with three tabular baselines under a chronological evaluation of the ULB–Worldline dataset. On the test period (52 frauds), the MLP achieved the best F1 (0.817) and MCC (0.823), Random Forest the best PR-AUC (0.765) and recall (0.750), and Logistic Regression the fewest false alarms (2) and the highest AUROC (0.977). Differences among the leading models amount to one or two frauds and a handful of false alerts, and are small relative to sampling uncertainty. The results therefore show that deep learning can offer a strong operating point, not that it replaces classical models; the preferred model depends on the relative cost of missed fraud, false alerts, and ranking quality. The main contribution is an auditable evaluation protocol: preserve chronology, select thresholds on validation data only, and report minority-class metrics with confusion counts. Because the data cover two days and use anonymized PCA features, the findings do not establish deployment readiness. Next steps are confidence intervals and multi-seed runs, validation on additional institutions and periods, and calibration, drift, cost-sensitive utility, latency, and human-review analysis.

References

[1] J. O. Awoyemi, A. O. Adetunmbi, and S. A. Oluwadare, "Credit card fraud detection using machine learning techniques: A comparative analysis," 2017 International Conference on Computing Networking and Informatics (ICCNI), 2017, doi: https://doi.org/10.1109/ICCNI.2017.8123782

[2] D. Prajapati, A. Tripathi, J. Mehta, K. Jhaveri, and V. Kelkar, "Credit Card Fraud Detection Using Machine Learning," 2021 International Conference on Advances in Computing, Communication, and Control (ICAC3), 2021, doi: https://doi.org/10.1109/ICAC353642.2021.9697227

[3] I. D. Mienye and N. Jere, "Deep Learning for Credit Card Fraud Detection: A Review of Algorithms, Challenges, and Solutions," IEEE Access, 2024, doi: https://doi.org/10.1109/ACCESS.2024.3426955

[4] A. A. Almazroi and N. Ayub, "Online Payment Fraud Detection Model Using Machine Learning Techniques," IEEE Access, 2023, doi: https://doi.org/10.1109/ACCESS.2023.3339226

[5] H. Shamsudin, U. K. Yusof, A. Jayalakshmi, and M. N. A. Khalid, "Combining oversampling and undersampling techniques for imbalanced classification: A comparative study using credit card fraudulent transaction dataset," 2020 IEEE 16th International Conference on Control & Automation (ICCA), 2020, doi: https://doi.org/10.1109/ICCA51439.2020.9264517

[6] M. Alamri and M. Ykhlef, "Hybrid Undersampling and Oversampling for Handling Imbalanced Credit Card Data," IEEE Access, 2024, doi: https://doi.org/10.1109/ACCESS.2024.3357091

[7] I. A. Mondal, M. E. Haque, A.-M. Hassan, and S. Shatabda, "Handling Imbalanced Data for Credit Card Fraud Detection," 2021 24th International Conference on Computer and Information Technology (ICCIT), 2021, doi: https://doi.org/10.1109/ICCIT54785.2021.9689866

[8] N. Prabha and S. Manimekalai, "Imbalanced data Classification in Credit Card Fraudulent Activities Detection using Multi-Class Neural Network," 2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS), 2022, doi: https://doi.org/10.1109/ICAIS53314.2022.9742878

[9] H. Aktar, M. A. Masud, N. J. Aunto, and S. N. Sakib, "Classification Using Random Forest on Imbalanced Credit Card Transaction Data," 2021 3rd International Conference on Sustainable Technologies for Industry 4.0 (STI), 2021, doi: https://doi.org/10.1109/STI53101.2021.9732553

[10] D. Dhiman, A. Bisht, A. Kumari, H. Anandaram, S. Saxena, and K. Joshi, "Online Fraud Detection using Machine Learning," 2023 International Conference on Artificial Intelligence and Smart Communication (AISC), 2023, doi: https://doi.org/10.1109/AISC56616.2023.10085493

[11] S. Lochan, H. V. Sumanth, A. Kodipalli, B. R. Rohini, T. Rao, and V. Pushpalatha, "Online Payment Fraud Detection Using Machine Learning," 2023 International Conference on Computational Intelligence for Information, Security and Communication Applications (CIISCA), 2023, doi: https://doi.org/10.1109/CIISCA59740.2023.00080

[12] H. M. M. H. V. Vidanelage, T. Tasnavijitvong, P. Suwimonsatein, and P. Meesad, "Study on Machine Learning Techniques with Conventional Tools for Payment Fraud Detection," 2019 11th International Conference on Information Technology and Electrical Engineering (ICITEE), 2019, doi: https://doi.org/10.1109/ICITEED.2019.8929952

[13] S. V. J. B. Gracia, J. G. Ponsam, S. Preetha, and J. G. K. Subhiksha, "Payment fraud detection using machine learning techniques," 2021 4th International Conference on Computing and Communications Technologies (ICCCT), 2021, doi: https://doi.org/10.1109/ICCCT53315.2021.9711887

[14] I. Almubark, "Advanced Credit Card Fraud Detection: An Ensemble Learning Using Random Under Sampling and Two-Stage Thresholding," IEEE

Access, 2024, doi: https://doi.org/10.1109/ACCESS.2024.3519335

[15] A. K. Yadav and V. Khanna, "Advanced Credit Card Fraud Detection Using Optimized Random Forest and Hybrid ELM with Ensemble Learning," 2025 7th International Conference on Artificial Intelligence and Speech Technology (AIST), 2025, doi: https://doi.org/10.1109/AIST68591.2025.11441551

[16] V. R. Adhegaonkar, A. R. Thakur, and N. Varghese, "Advancing Credit Card Fraud Detection Through Explainable Machine Learning Methods," 2024 2nd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT), 2024, doi: https://doi.org/10.1109/IDCIoT59759.2024.10467999

[17] D. Varmedja, M. Karanovic, S. Sladojevic, M. Arsenovic, and A. Anderla, "Credit Card Fraud Detection - Machine Learning methods," 2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH), 2019, doi: https://doi.org/10.1109/INFOTEH.2019.8717766

[18] K. Hemanth, K. S. Virat, M. D. Rohith, K. V. P. Reddy, A. S. Selv, and S. Pandi S, "Credit Card Fraud Detection using Machine Learning Methods," 2025 Emerging Technologies for Intelligent Systems (ETIS), 2025, doi: https://doi.org/10.1109/ETIS64005.2025.10961927

[19] H. A. Atia, M. Aboul-Ela, C. A. Reyad, and N. A. Awad, "Online Payments Fraud Detection Using Machine Learning Techniques," 2024 Intelligent Methods, Systems, and Applications (IMSA), 2024, doi: https://doi.org/10.1109/IMSA61967.2024.10652834

[20] U. Siddaiah, P. Anjaneyulu, Y. Haritha, and M. Ramesh, "Fraud Detection in Online Payments using Machine Learning Techniques," 2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS), 2023, doi: https://doi.org/10.1109/ICICCS56967.2023.10142404

[21] S. Lu, "Graph Neural Network Model in Financial Fraud Detection," 2025 2nd International Conference on Intelligent Computing and Robotics (ICICR), 2025, doi: https://doi.org/10.1109/ICICR65456.2025.00176

[22] A. Kesharwani and P. Shukla, "FFDM-GNN: A Financial Fraud Detection Model using Graph Neural Network," 2024 International Conference on Computing, Sciences and Communications (ICCSC), 2024, doi: https://doi.org/10.1109/ICCSC62048.2024.10830438

[23] A. C. Hiremath, A. Arya, L. N. Sriranga, K. V. S. R. Reddy, and M. Nikhil, "Ensemble of Graph Neural Networks for Enhanced Financial Fraud Detection," 2024 IEEE 9th International Conference for Convergence in Technology (I2CT), 2024, doi: https://doi.org/10.1109/I2CT61223.2024.10543898

[24] Y. Cui, X. Han, J. Chen, X. Zhang, J. Yang, and X. Zhang, "FraudGNN-RL: A Graph Neural Network With Reinforcement Learning for Adaptive Financial Fraud Detection," IEEE Open Journal of the Computer Society, vol. 6, pp. 426–437, 2025, doi: https://doi.org/10.1109/OJCS.2025.3543450

[25] T. Haripriya, S. Chandramouli, M. Kavitha, M. D. Parne, D. Srinivas, and B. K. Bala, "Machine Learning Algorithms for Dynamic Financial Management and Fraud Detection: Implementing Anomaly Detection, and Risk Assessment Techniques," 2025 International Conference on Intelligent Systems and Computational Networks (ICISCN), 2025, doi: https://doi.org/10.1109/ICISCN64258.2025.10934386

[26] B. Kumar, S. K. Gupta, and M. Patnaik, "Machine Learning-Powered Fraud Detection & Prevention: A Comprehensive Implementation," 2024 2nd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT), 2024, doi: https://doi.org/10.1109/IDCIoT59759.2024.10467714

[27] G. Alessi and M. Fugini, "Real-Time Fraud Detection Using Machine Learning," 2025 33rd International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE), 2025, doi: https://doi.org/10.1109/WETICE67341.2025.11092218

[28] S. Mahadik, P. Chopra, N. Kassetty, V. Ravalji, O. Goel, and J. K. Gupta, "Developing Machine Learning Models for Real-Time Fraud Detection in Online Transactions," 2025 International Conference on Networks and Cryptology (NETCRYPT), 2025, doi: https://doi.org/10.1109/NETCRYPT65877.2025.11102173

[29] G. Liu, J. Tang, Y. Tian, and J. Wang, "Graph Neural Network for Credit Card Fraud Detection," 2021 International Conference on Cyber-Physical Social Intelligence (ICCSI), 2021, doi: https://doi.org/10.1109/ICCSI53130.2021.9736204

[30] H. Zhu, M. Zhou, G. Liu, Y. Xie, S. Liu, and C. Guo, "NUS: Noisy-Sample-Removed Undersampling Scheme for Imbalanced Classification and Application to Credit Card Fraud Detection," IEEE Transactions on Computational Social Systems, vol. 11, no. 2, 2024, doi: https://doi.org/10.1109/TCSS.2023.3243925

[31] Machine Learning Group–ULB and Worldline, "Credit Card Fraud Detection," Kaggle dataset. https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud