International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 6, Nov-Dec-2026, ISSN (Online): 2395-566X

Machine Learning-Based Credit Card Fraud Detection (A Thematic Review)

Sukhpreet Kaur[1] Assistant Professor Ms. Jaspreet Kaur[2]
[1]MCA students Guru Kashi University Talwandi Sabo Bathinda Punjab India
[2]Assistant Professor Faculty of computing Guru Kashi University Talwandi Sabo Bathinda Punjab India{ Rudramurthy.Cbs,

Abstract- Credit card use has expanded rapidly, and with it the opportunities for fraud that is hard to distinguish from legitimate spending. Machine learning has become the main research response, yet published results are difficult to compare and harder to translate into operational practice. This review synthesises 22 scholarly publications drawn from the reference base of a master's dissertation on the topic and reorganizes them into thematic strands: classifier benchmarking, feature engineering and hybrid designs, sequence and generative approaches, data availability and class imbalance, and evaluation, explainability and concept drift. The method is a reconstructed, non-systematic thematic synthesis; no new database search or screening was carried out. The reviewed work indicates that tree-based ensembles are the most frequently applied and often the strongest supervised models, that behaviorally informed features and hybrid designs add value beyond the choice of algorithm, and that accuracy is an inadequate measure when fraud makes up a small fraction of one per cent of transactions. The evidence base, however, rests heavily on one anonymized public dataset and on confidential industrial data, which limits reproducibility and generalization. Five gaps stand out: scarce real data, non-comparable evaluation protocols, concept drift and delayed labels, limited explainability, and the absence of realtime and local validation, including for the Indian market. Research directions are proposed for each gap, and standardized temporal evaluation is argued to be the most immediate need.

Keywords: credit card fraud; fraud detection; machine learning; class imbalance; concept drift; ensemble learning; explainability

I. INTRODUCTION

Credit card payments are checked and approved by the issuing bank. The bank checks the available credit limit and transaction details before approving a payment. The Reserve Bank of India (RBI) provides data about credit card transactions, including online and POS payments. Although credit cards make payments easy and convenient, they also involve security risks. Fraudulent transactions can happen within seconds. Therefore, effective fraud detection systems are important to protect users and prevent financial losses. The use of credit cards is increasing rapidly in India and other countries.

Along with this growth, credit card fraud has also become a major security problem. One of the biggest challenges for researchers is the lack of large and publicly available real-world transaction data. Banks cannot easily share customer details because of privacy and security reasons.

1. Background and importance Credit cards have become one of the most visible instruments of digital payment in India. Reserve Bank of India (RBI) statistics put outstanding cards at about 124.05 million in August 2026, up from about 112.34 million a year earlier, with monthly spending of roughly ₹2.02 trillion spread across point-ofsale and online channels (RBI, n.d.). A payment system of this size has to authorize each transaction within seconds, and fraudsters exploit that speed. The RBI's annual report for 2024-25 records that card and internet-related frauds account for the largest share of reported bank frauds by number, although loanrelated frauds account for most of the value (RBI, 2025). Card fraud takes several forms, including lost or stolen cards, counterfeit or skimmed cards, cardnotpresent purchases with stolen credentials, application fraud, account takeover and phishing-led credential theft (Kaur, 2026). What makes detection hard is that a fraudulent transaction can look ordinary: the amount may be typical, the card genuine, and the authentication successful. Fraud therefore rarely reveals itself in a single transaction attribute, and the problem is better understood as one recognizing deviation from expected behavior over time.

2. Research landscape and motivation Early work on statistical fraud detection set out the distinction between supervised methods, which learn from labelled fraud, and unsupervised methods, which flag unusual behavior (Bolton & Hand, 2002). Rule-based systems remain common in practice, but they detect only patterns that someone has already documented. Machine learning offers a way of learning such patterns from historical data, and the card fraud literature has grown to include classical classifiers, boosted trees, sequence models, and generative methods. Several surveys now map this territory (Boush et al., 2023; Cherif et al., 2023; Phua et al., 2010).

© 2026 IJSRET

The difficulty for a newcomer, and arguably for the field, is that this literature is fragmented. Studies use different datasets, different preprocessing and different metrics, so that the headline figure in one paper cannot be set beside the headline figure in another. The motivation for this review is to bring order to that fragmentation by reading the literature thematically rather than study by study.

3. Objectives, scope and organization

The review has four objectives: to synthesize how machine learning has been applied to card fraud detection; to compare the datasets, techniques and evaluation practices reported; to identify the gaps that limit the reliability of current findings; and to propose research directions that address them. It is confined to credit card fraud and to supervised, unsupervised, hybrid and deep learning approaches. Broader digital payment fraud, system architecture, and commercial products are outside its scope. Section 2 describes the review method and its limits. Section 3 presents the thematic review; Section 4 offers an integrated critical discussion; Sections 5 and 6 set out research gaps and future directions, and Section 7 concludes.

II. REVIEW METHODOLOGY

This is a narrative thematic review. It is not a systematic review, and the method described here is a reconstructed one. The source dissertation (Kaur, 2026) states that it follows a secondary research approach, drawing on published papers, institutional reports and public datasets, and comparing selected studies by objective, dataset, algorithm, evaluation method and conclusion.

It does not report databases searched, search strings, date limits or formal inclusion and exclusion criteria. No database search, record screening, or quality appraisal was performed for the present paper, and none is claimed.

The corpus is the dissertation reference base: 22 scholarly publications, together with the ULB benchmark dataset record (Machine Learning Group - ULB, n.d.) and two RBI sources used for context. Selection was therefore purposive and inherited. The scholarly items were assigned to thematic strands according to their main contribution, as shown in Table 1, and some appear under more than one strand because their contribution cuts across themes. Within each strand, the studies were compared by data source, technique, evaluation measure and reported outcome, and agreements and disagreements were noted.

A quantitative synthesis was not attempted because the reported metrics are computed on different data under different protocols and cannot be pooled.

Two caveats follow. First, the corpus is not exhaustive, so absences noted in this review (for example, of a particular kind of study) are absences from this corpus and not necessarily from the field. Second, the descriptions of individual studies follow the dissertation's account of them and the bibliographic record; the primary papers were not independently re-read in full for this version, and any figure attributed to a study should be checked against the original before submission.

Table 1 Thematic Classification of the Reviewed Corpus

Thematic strand	Principal sources
Fraud context and surveys	Bolton and Hand (2002); Phua et al. (2010); Cherif et al. (2023); Boush et al. (2023); RBI (n.d., 2025)
Classifier benchmarking	Bhattacharyya et al. (2011); Dorna Dula and Geetha (2019); Alarfaj et al. (2022); Taha and Malebary (2020); foundations in Breiman (2001) and Chen and Guestrin (2016)
Features, hybrid and ensemble designs	Whitrow et al. (2009); Bahnsen et al. (2016); Randhawa et al. (2018); Carcillo et al. (2021); Zhang et al. (2021); Paldino et al. (2024)
Sequence and generative approaches	Jurgovsky et al. (2018); Fiore et al. (2019)
Data availability and class imbalance	Dal Pozzolo et al. (2015, 2018); Chawla et al. (2002); Machine Learning Group - ULB (n.d.); Paldino et al. (2024)
Evaluation, explainability and drift	Lundberg and Lee (2017); Dal Pozzolo et al. (2018); Paldino et al. (2024); Bahnsen et al. (2016); Btoush et al. (2023)

© 2026 IJSRET

III. THEMATIC LITERATURE REVIEW

1. Benchmarking classifiers

The most common form of study compares a set of off-the-shelf classifiers on a single dataset.

Bhattacharyya et al. (2011) is an early and influential example: using real data from an international credit card operation, they compared Logistic Regression, Support Vector Machines and Random Forests and found the Random Forests to perform better overall. Dornadula and Geetha (2019) widened the comparison to include anomaly-detection algorithms and applied SMOTE to address imbalance. Alarfaj et al. (2022) compared machine learning and deep learning models on the European cardholder benchmark and also ran experiments on balanced data aimed at reducing false negatives; in the dissertation's account, their proposed deep learning model outperformed the models it was compared with. Taha and Malebary (2020) reported an optimized light gradient boosting approach.

The surveys give a broader picture. Btoush et al. (2023), reviewing 181 journal articles from 2019 to 2021, reported that Random Forest was the most commonly applied model, appearing in 74 articles, followed in prominence by Logistic Regression and Support Vector Machine, and that supervised techniques dominated the field. Cherif et al. (2023), reviewing 40 studies published between 2015 and 2021, concluded that deep learning had been investigated only to a limited extent. Taken together, these sources suggest a rough convergence on tree-based ensembles as the default strong learner. That is plausible on methodological grounds, since Random Forests combine many decision trees and are comparatively robust to overfitting (Breiman, 2001), while gradient boosting scales well on structured data (Chen & Gestring, 2016). The explanation is, however, an interpretation; the studies themselves mostly report outcomes rather than test mechanisms.

The apparent disagreement between Alarfaj et al. (2022), where a deep model leads, and the broader ensemble-centered picture should be read with care. Authors who propose a model tend to report it as superior to the baselines they selected, and the comparison sets, tuning effort and data differ from paper to paper. The two positions are not necessarily contradictory; they are simply not measured on common ground.

2. Features, hybrid designs and ensembles

A second strand moves the contribution away from the choice of classifier and towards the representation of the data. Whitrow et al. (2009) showed the value of aggregating a cardholder's recent transactions rather than examining each transaction in isolation. Bahnsen et al. (2016) extended this with aggregation and periodic time-based features and evaluated them using a cost-based measure, which ties model assessment to the financial consequences of errors. Zhang et al. (2021) proposed a homogeneity-oriented behavior analysis (HOBA) feature-engineering method combined with a deep learning classifier.

Hybrid designs take a similar view from the modeling side. Carcillo et al. (2021) computed unsupervised outlier scores at several levels of granularity and supplied them as inputs to a supervised classifier, and reported improved detection on a real, annotated card fraud dataset. Randhawa et al. (2018) combined standard classifiers with AdaBoost and majority voting and used the Matthews correlation coefficient for evaluation. Paldino et al. (2024) examined diversity in ensembles and proposed ensembles that preserve past concepts so that a model can adapt more quickly when behavior changes.

These studies share an assumption: that a single transaction carries limited evidence, and that behavioral context carries more. They differ in where the context is injected, whether through hand-built aggregates, unsupervised scores or ensemble structure. The studies by Carcillo et al. (2021) and Paldino et al. (2024) used real industrial data, which strengthens their relevance but, because that data cannot be shared, limits independent verification. One consequence of this strand deserves emphasis. Aggregation strategies depend on being able to link transactions to a cardholder, whereas the public benchmark exposes only anonymized components together with Time and Amount (Machine Learning Group - ULB, n.d.). It follows, as an inference from the data description and not a finding of any reviewed study, that the most behaviorally informed methods cannot be straightforwardly reproduced on the most widely used public data.

3. Sequence and generative approaches Deep learning enters the literature through two distinct routes. Jurgovsky et al. (2018) formulated card fraud detection as a sequence classification problem and compared an LSTM network with a Random Forest using aggregated features, showing that sequence models can capture behavioral information that isolated transactions miss. Fiore et al. (2019) used generative adversarial networks (GANs) to produce additional minority-class examples and reported that a classifier trained on the augmented data performed better than one trained on the original data.

© 2026 IJSRET

The two studies address different problems. The first changes how the data are represented, and the second changes how scarce the positive class is. Their fit with the earlier strands is also different: a sequence model can be seen as an alternative to hand-built aggregation, whereas GAN-based augmentation competes with resampling techniques such as SMOTE (Chawla et al., 2002). In the studies as described in the dissertation, there is no head-to-head comparison between these alternatives, so claims that one is preferable remain untested within this corpus. Deep models also generally demand more data, more computation and more careful design, and they are harder to interpret, which matters to the reasons discussed in Section 3.5.

4. Data availability and class imbalance

Data availability shapes almost everything else in the field. Transaction records contain sensitive financial and personal information, so institutions rarely release them. The most widely used public benchmark is the dataset from the Machine Learning Group of Universite Libre de Bruxelles: 284,807 transactions by European cardholders over two days in September 2013, of which 492 (0.172%) are fraudulent. Most input variables are principal components produced by PCA for confidentiality, and only Time and Amount are left untransformed (Dal Pozzolo et al., 2015; Machine Learning Group - ULB, n.d.). At the other extreme, Dal Pozzolo et al. (2018) tested their learning strategy on more than 75 million e-commerce card transactions, and Paldino et al. (2024) used extracts of two real datasets of more than 30 million and 50 million transactions supplied by an industrial partner.

The result is a bifurcated evidence base. Public results are reproducible but rest on a small, anonymized and short-span sample; industrial results are realistic but cannot be reproduced. The two-day span of the benchmark also means that it offers little scope for studying to change over time, again an inference from the dataset description.

Imbalance is handled in several ways: class weighting, under sampling (Dal Pozzolo et al., 2015), oversampling with SMOTE (Chawla et al., 2002) and GAN-based generation of synthetic frauds (Fiore et al., 2019). Each alters the effective class distribution seen by the learner.

Dal Pozzolo et al. (2015) address the consequence directly, as their title indicates, by calibrating probabilities after under sampling. The point matters because a fraud score is often used as a probability or ranking, and a model trained on rebalanced data will not report probabilities that reflect true fraud prevalence unless corrected.

5 Evaluation, explainability and concept drift

The reasoning behind the demand for better metrics is straightforward. If 99% of transactions are legitimate, a classifier that labels everything legitimate achieves 99% accuracy while detecting no fraud. The literature therefore emphasizes precision, recall, F1-score, ROC-AUC and confusion matrices, with some studies adding the Matthews correlation coefficient (Randhawa et al., 2018) or cost-based measures (Bahnsen et al., 2016); Dal Pozzolo et al. (2018) discuss which measures are appropriate for fraud detection. Precision-recall measures are often favored for extreme imbalance because they concentrate on the minority class. Despite this, Btoush et al. (2023) noted that 24 of the 181 studies they reviewed reported accuracy as their only performance metric, so the recommended practice is far from universal.

Explainability is the least developed strand. Shapley Additive explanations (SHAP) assign each feature a contribution to an individual prediction (Lundberg & Lee, 2017) and can be applied to tree-based models such as XGBoost (Chen & Guestrin, 2016). For card issuers, a score without a reason is of limited use, since analysts and customers may need to know why a transaction was flagged. Yet the corpus contains no empirical study that applies such tools to card fraud models; the link is a natural one, as the dissertation puts it, but still a developing direction.

Concept drift and delayed labels to complete the picture. Dal Pozzolo et al. (2018) describe the operating conditions of a real fraud detection system, including class imbalance, concept drift, and verification latency. The delay before the true label of an alert becomes known. Paldino et al. (2024) show that ensembles preserving past concepts can adapt faster when the data distribution changes. Both papers treat the problem as a continuing process rather than a one-off classification exercise, which contrasts with the many offline studies evaluated on a fixed historical dataset.

Table 2 Summary of Selected Studies in the Reviewed Corpus

Study	Approach	Data (as reported in the dissertation)	Main reported finding
Bhattacharyya et al. (2011)	Logistic Regression, SVM, Random Forest	Real international credit card data	Random Forests performed better overall

© 2026 IJSRET

Dorna Dula & Geetha (2019)	Supervised and anomalydetection algorithms; SMOTE	Not stated in the dissertation	Compared several algorithms under imbalance handling
Alarfaj et al. (2022)	Machine learning and deep learning comparison; balanced-data experiments	European cardholder benchmark	Proposed deep model outperformed comparators; balancing reduced false negatives
Randhawa et al. (2018)	Classifiers with AdaBoost and majority voting	Not stated in the dissertation	Ensemble route, evaluated with the Matthews correlation coefficient
Carcillo et al. (2021)	Unsupervised outlier scores feeding a supervised classifier	Real, annotated card fraud dataset	Combination improved detection accuracy
Jurgovsky et al. (2018)	LSTM sequence model versus Random Forest with aggregated features	Not stated in the dissertation	Sequence models capture additional behavioral information
Fiore et al. (2019)	GAN-generated minorityclass examples	Real credit card transactions	Classifier trained on augmented data outperformed the original
Dal Pozzolo et al. (2018)	Realistic fraud detection model and learning strategy	More than 75 million ecommerce transactions	Addresses imbalance, concept drift and verification latency
Paldino et al. (2024)	Diversity-based ensembles preserving past concepts	Two real datasets of over 30 million and 50 million transactions	Faster adaptation when behavior changes

4. CRITICAL ANALYSIS AND DISCUSSION

Read together, the strands show a field whose center of gravity has moved. The earliest comparative work asked which algorithm performs best. More recent contributions ask under what conditions a model has to operate: with what behavioral context, against what class distribution, with what delay in feedback and with what cost of error. Whitrow et al. (2009), Bahnsen et al. (2016) and Zhang et al. (2021) relocate the problem to feature construction, Jurgovsky et al. (2018) to sequence structure, and Dal Pozzolo et al. (2018) and Paldino et al. (2024) to the dynamics of a deployed system. This is the most important trend in the corpus, and it implies that further gains from swapping one classifier for another are likely to be smaller than gains from better problem formulation.

Methodologically, the principal weakness is the comparability of results. The dissertation illustrates the point well: a model reporting 98% accuracy on one dataset cannot be judged better than one reporting 93% on another (Kaur, 2026). Three sources of non-comparability recur: the dataset, the preprocessing and feature set, and the metric. A fourth, more subtle, risk is that resampling applied before the data are split into training and test sets can inflate results; this is a general caution about the technique and not a finding about any study reviewed here, but it underlines why evaluation protocols should be reported in detail. The convergence on Random Forest and boosting is probably the sturdiest empirical regularity, because it appears across surveys that cover different periods and many studies. Even so, it is a regularity about popularity and relative performance in reported comparisons, and not a demonstration of general superiority.

© 2026 IJSRET

The evidence on deep learning is genuinely mixed. Cherif et al. (2023) found limited investigation, Alarfaj et al. (2022) reported a leading deep model, and Jurgovsky et al. (2018) and Fiore et al. (2019) show specific settings in which sequence or generative methods add value. A reasonable reading is that deep methods are promising where behavioral sequences or severe scarcity of positives are central, and that their benefit over well-engineered ensembles on tabular features has not been established. The costs in data, computation, and interpretability are real.

The practical implications follow this. For an issuer, model choice is not only a matter of detection rate; it also involves the cost of false alarms, the ability to explain decisions, and the capacity to update fraud changes. For researchers, the implication is that claims should be accompanied by temporal validation, full reporting of imbalance treatment, and metrics that reflect the minority class. Finally, the geographic dimension should not be overlooked. The motivating statistics in this review concern India, whereas the reproducible evidence concerns European cardholders in 2013. The dissertation is explicit that this limits confidence in generalizing to the Indian market (Kaur, 2026), and nothing in the reviewed corpus resolves that limitation.

V. RESEARCH GAPS

The gaps below are drawn from the dissertation's own list and from the reviewed literature and are organized by type in Table 3. Some are explicit in the sources, such as data scarcity and drift; others, flagged as such, are observations arising from the synthesis.

Theoretically, the corpus offers few explanations of why particular models succeed. Claims about the strength of tree ensembles are largely empirical, and the interaction between imbalance, drift and delayed labels is described by Dal Pozzolo et al. (2018) but not yet matched by a widely used evaluation framework. This is an observation of this review. Methodologically, the lack of shared protocols is the central obstacle: different datasets, preprocessing and metrics make reported accuracy values incomparable, and a minority of studies still report accuracy alone (Btoush et al., 2023). Empirically, real card data are scarce, the public benchmark is small and short-lived, and there is little evidence from the Indian setting. Practically, explainability, real-time operation, and regulatory acceptability remain largely unaddressed, because most studies are offline for evaluations on historical data. Among emerging opportunities, the sharing of data in privacy-preserving forms stands out, since the supply of realistic data limits every other line of work.

Table 3
Identified Research Gaps and Their Significance

Gap type	Gap	Why it matters	Basis
Theoretical	Limited explanation of why ensembles perform well; no common framework combining imbalance, drift and label delay	Without explanation, model choice rests on habit and reported results	Dal Pozzolo et al. (2018); observation of this review
Methodological	Non-comparable datasets, preprocessing and metrics; accuracy-only reporting; mostly offline evaluation	Prevents cumulative progress and direct comparison	Btoush et al. (2023); Kaur (2026)
Empirical	Scarcity of real, shareable data; a single short-span public benchmark; little evidence from Indian card data	Results may not transfer to real or local conditions	Dal Pozzolo et al. (2015); Kaur (2026)
Practical	Limited explainability and realtime evaluation; concept drift and verification latency	Issuers need reasons, speed and sustained accuracy	Lundberg & Lee (2017); Dal Pozzolo et al. (2018); Paldino et al. (2024)
Emerging	Privacy-preserving data sharing and synthetic data	Would widen access to realistic data	Observation; Fiore et al. (2019) for generative augmentation

Emerging Privacy-preserving data sharing and synthetic data Would widen access to realistic data Observation; Fiore et al. (2019) for generative augmentation

© 2026 IJSRET

1. Future Research Directions

The directions below are proposals arising from the gaps above and not findings of the reviewed studies. Technologies mentioned beyond the corpus are named as candidate approaches, not as established solutions.

A priority is a common evaluation protocol. A useful question is how the ranking of Random Forest, boosted trees, sequence models and hybrid designs changes when all are evaluated on the same data under strictly temporal train and test splits, with precision-recall and cost-based measures reported alongside conventional ones. Where only the public benchmark is available, results should be framed as relative comparisons and the limitations of its short time span made explicit. Related questions are how much behavioral feature engineering adds when the classifier is held constant, and whether the gain from hybrid and sequence designs survives a fair comparison with strong, well-tuned ensembles.

The second line concerns imbalance and calibration. Controlled studies could compare class weighting, under sampling, SMOTE and GAN-based augmentation on the same data, assessing not only detection performance but the calibration of scores and robustness under drift, which the present corpus does not compare directly. A third line concerns operation: evaluating models under delayed labels, in streaming settings and within latency limits would bring academic evaluation closer to authorization conditions. Explainability can be studied empirically by applying SHAP or similar methods to fraud models and asking whether the explanations are stable and useful to analysts, a question that calls for collaboration with practitioners and for human-centered evaluation.

Data access underpins all of these. Collaborations between universities, issuers and regulators could yield access under controlled conditions, and privacy-preserving approaches such as synthetic data generation or federated learning merit examination as ways of sharing knowledge without sharing records. Studies using Indian card data, separated by channel (point-of-sale and online), would test whether findings from European data transfer. The topic is inherently interdisciplinary, drawing on payment regulation, behavioral economics for cost-sensitive decision-making, and human-computer interaction for analyst-facing tools.

VI. CONCLUSION

The literature on machine learning for credit card fraud detection is mature enough to show clear regularities and young enough to leave basic questions unanswered. Tree-based ensembles are the most frequently reported strong supervised models; richer behavioral features, hybrid designs, sequence models and generative augmentation each add something in specific settings; and accuracy alone is an unreliable guide when fraud is rare.

The most consequential finding of this synthesis concerns the evidence itself. It rests on one anonymized public benchmark and private industrial data, so models can be ranked within studies far more securely than across them or against real-world fraud.

The gaps identified here that namely scarce real data, non-comparable evaluation, concept drift and label delay, limited explainability and the lack of real-time and local validation, are connected, and progress on one tends to depend on progress on the others. The contribution of this review is to organize a fragmented body of work into a thematic structure, to separate what the studies report from what this synthesis infers, and to propose a research agenda in which standardized temporal evaluation comes first. Because the review is built on a dissertation reference base and not on a fresh systematic search, it should be read as a structured starting point, and a systematic update would be a worthwhile extension.

REFERENCES

1. Alarfaj, F. K., Malik, I., Khan, H. U., AL Musallam, N., Ramzan, M., & Ahmed, M. (2022). Credit card fraud detection using state-of-the-art machine learning and deep learning algorithms. IEEE Access, 10, 39700–39715. https://doi.org/10.1109/ACCESS.2022.3166891
2. Bahnsen, A. C., Aouad, D., Stojanovic, A., & Ottenstein, B. (2016). Feature engineering strategies for credit card fraud detection. Expert Systems with Applications, 51, 134–142.
3. Bhattacharyya, S., Jha, S., Thara Kunnel, K., & Westland, J. C. (2011). Data mining for credit card fraud: A comparative study. Decision Support Systems, 50(3), 602–613.
4. Bolton, R. J., & Hand, D. J. (2002). Statistical fraud detection: A review. Statistical Science, 17(3), 235–255.
5. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.
6. https://doi.org/10.1023/A:1010933404324
7. Btoush, E. A. L. M., Zhou, X., Gururajan, R., Chan, K. C., Genrich, R., & Sankaran, P. (2023). A systematic review of literature on credit card cyber fraud detection using machine and deep learning. Peer Computer Science, 9, Article e1278. https://doi.org/10.7717/peerj-cs.1278

© 2026 IJSRET

8. Carcillo, F., Le Borgne, Y.-A., Caelen, O., Kessaci, Y., Obel, F., & Bontempi, G. (2021). Combining unsupervised and supervised learning in credit card fraud detection. Information Sciences,
9. 557, 317–331. https://doi.org/10.1016/j.ins.2019.05.042
10. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kielmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357. https://doi.org/10.1613/jair.953
11. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the
12. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining
13. (pp. 785–794). ACM. https://doi.org/10.1145/2939672.2939785
14. 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. http://www.sciencedirect.com/science/article/pii/S1319157822004062
15. Dal Pozzolo, A., Borachio, G., Caelen, O., Aleppo, C., & Bontempi, G. (2018). Credit card fraud detection: A realistic modeling and a novel learning strategy. IEEE Transactions on Neural Networks and Learning Systems, 29(8), 3784–3797. https://doi.org/10.1109/TNNLS.2017.2736643
16. Dal Pozzolo, A., Caelen, O., Johnson, R. A., & Bontempi, G. (2015). Calibrating probability under sampling for unbalanced classification. In 2015 IEEE Symposium Series on Computational Intelligence (SSCI) (pp. 159–166). IEEE. https://doi.org/10.1109/SSCI.2015.33
17. Dorna Dula, V. N., & Geetha, S. (2019). Credit card fraud detection using machine learning algorithms. Procedia Computer Science, 165, 631–641.
18. Fiore, U., De Santis, A., Perla, F., Zanetti, P., & Palmieri, F. (2019). Using generative adversarial networks to improve classification effectiveness in credit card fraud detection. Information Sciences, 479, 448–455.
19. Jirkovsky, J., Granitzer, M., Ziegler, K., Calabretta, S., Portier, P.-E., He-Galeton, L., & Caelen, O. (2018). Sequence classification for credit card fraud detection. Expert Systems with Applications, 100, 234–245.
20. Kaur, S. (2026). Machine learning-based fraud detection in credit card transactions [master's dissertation, Guru Kashi University].
21. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems 30 (pp. 4765–4774).
22. Machine Learning Group - ULB. (n.d.). Credit card fraud detection [Data set]. Kaggle.
23. https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud
24. Paldino, G. M., Lebichot, B., Le Borgne, Y.-A., Siblini, W., Oblé, F., Boracchi, G., & Bontempi, G.
25. (2024). The role of diversity and ensemble learning in credit card fraud detection. Advances in
26. Data Analysis and Classification, 18(1), 193–217. https://doi.org/10.1007/s11634–022– 00515–5
27. Phua, C., Lee, V., Smith, K., & Gayler, R. (2010). A comprehensive survey of data mining-based fraud detection research. arXiv. https://arxiv.org/abs/1009.6119
28. Randhawa, K., Loo, C. K., Seera, M., Lim, C. P., & Nandi, A. K. (2018). Credit card fraud detection using AdaBoost and majority voting. IEEE Access, 6, 14277–14284.
29. Reserve Bank of India. (n.d.). Bank-wise ATM/POS/card statistics. Retrieved October 15, 2026, from https://rbi.org.in/Scripts/ATMView.aspx
30. Reserve Bank of India. (2025). Annual report 2024–25. Retrieved October 15, 2026, from https://www.rbi.org.in/Scripts/AnnualReportPublications.aspx
31. Taha, A. A., & Malebary, S. J. (2020). An intelligent approach to credit card fraud detection using an optimized light gradient boosting machine. IEEE Access, 8, 25579–25587.
32. Whitrow, C., Hand, D. J., Juszczak, P., Weston, D., & Adams, N. M. (2009). Transaction aggregation as a strategy for credit card fraud detection. Data Mining and Knowledge Discovery, 18(1), 30–55.
33. Zhang, X., Han, Y., Xu, W., & Wang, Q. (2021). HOBA: A novel feature engineering methodology for credit card fraud detection with a deep learning architecture. Information Sciences, 557, 302– 316.

© 2026 IJSRET