© AUG 2026 | IRE Journals | Volume 10 Issue 2 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I2-1722685

A Comparative Evaluation of Machine Learning Models for Credit Card Fraud Detection in E-Commerce

ZIAKEGHA LUCKY TONBRAPAGHA[1], ZIAKEGHA CLEMENT EBIMOBOWEI[2]
[1]*Department of Computer Science, Niger Delta University, Wilberforce Island, Bayelsa State, Nigeria*
[2]*National Open University, Nigeria*

Abstract- ***The rapid growth of e-commerce and electronic payment services has increased the convenience of digital transactions while also creating more opportunities for fraudulent activity. Conventional fraud detection approaches based on manual review and fixed rules can struggle with large transaction volumes, changing fraud patterns, and false-positive alerts. This study comparatively evaluates four supervised machine learning algorithms-Logistic Regression, K-Nearest Neighbors (KNN), Decision Tree, and Support Vector Machine (SVM) for credit card fraud detection. The study used a labelled transaction dataset and followed an iterative Agile-based development process involving data preparation, model training, testing, and performance evaluation. The models were assessed using accuracy, precision, recall, F1-score, receiver operating characteristic area under the curve (ROC-AUC), and training time. The reported experimental results show that Logistic Regression achieved the highest accuracy (63.33%), precision (67.61%), and ROC-AUC (0.6855), while KNN produced the highest recall (67.50%) and F1-score (0.6506). SVM produced moderate results but required the longest training time of 0.0345 seconds. Decision Tree had the shortest training time at 0.0058 seconds but recorded the weakest overall predictive performance, including 54.00% accuracy and 40.00% recall. The findings demonstrate that model selection for fraud detection should not be based on accuracy alone because missing fraudulent transactions can have serious financial and operational consequences. Logistic Regression provided the strongest overall balance among the evaluated models, although the moderate performance of all four algorithms indicates that larger and more representative datasets, improved feature engineering, class-imbalance treatment, and hyperparameter optimization are required before practical deployment.***

Keywords: ***credit card fraud, e-commerce, machine learning, fraud detection, supervised learning***

I. INTRODUCTION

Electronic commerce has changed the way consumers purchase goods and services and how businesses reach markets. Online stores, digital payment platforms, mobile applications, and other electronic channels allow transactions to take place with little geographical or temporal restriction. The expansion of these channels has created significant benefits for consumers and businesses, but it has also increased exposure to fraud. Credit card fraud, identity theft, account takeover, phishing-related payment abuse, and other forms of online fraud can result in direct financial losses and can weaken consumer confidence in digital commerce [1–4].

Fraud detection is particularly difficult because fraudulent transactions are relatively rare compared with legitimate transactions and because fraudulent behaviour changes over time. Traditional approaches have commonly relied on manual investigation and predefined rules. Although these approaches remain useful, they may generate large numbers of alerts and may fail when fraudsters change their behaviour in ways not covered by existing rules. Previous research has therefore increasingly considered data mining and machine learning techniques as ways of learning patterns from transaction histories and supporting automated classification [2, 5, 6].

Supervised machine learning is particularly relevant when historical transactions are labelled as legitimate or fraudulent. Algorithms can learn relationships between transaction attributes and the known class labels and then apply those learned relationships to previously unseen transactions. Different algorithms, however, make different assumptions about the data and provide different trade-offs between predictive performance, interpretability, computational cost, and

sensitivity to class imbalance [3, 7–10]. Consequently, comparing several models under the same experimental conditions is useful for determining which approach provides the most appropriate balance for a particular fraud-detection problem.

This study therefore evaluates Logistic Regression, KNN, Decision Tree, and SVM using a common transaction dataset and a common set of performance measures. The study is motivated by the need to move beyond accuracy as the sole indicator of fraud-detection quality and to consider metrics that reflect the consequences of false positives and false negatives. The paper presents the methodology, experimental results, comparative discussion, and implications of the findings for machine-learning-based fraud detection in e-commerce.

II. RELATED LITERATURE

2.1 Fraud Detection in Digital Transactions

Fraud detection has traditionally involved statistical analysis, expert judgement, manual investigation, and rule-based monitoring. Bolton and Hand [2] describe statistical fraud detection as a field concerned with identifying observations that depart from expected patterns. In financial systems, this objective is complicated by the fact that legitimate customer behaviour is itself diverse, while fraudulent behaviour may deliberately imitate legitimate activity. Ngai et al. [5] similarly show that data mining techniques have become important in financial fraud detection because transaction data can contain patterns that are difficult to identify through manual inspection alone.

The practical objective of a fraud detection system is not simply to classify transactions correctly. A useful system should identify as many fraudulent transactions as possible while limiting the number of legitimate transactions incorrectly blocked or investigated. Bhattacharyya et al. [3] emphasize the value of comparative evaluation when choosing fraud-detection methods, while Bahnsen et al. [6] demonstrate the importance of considering the cost associated with different types of classification errors.

2.2 Supervised Machine Learning for Fraud Detection

Supervised learning uses labelled examples to learn a mapping between input variables and a target class. In credit card fraud detection, the target commonly represents legitimate and fraudulent transactions. The literature includes statistical classifiers, tree-based models, distance-based classifiers, margin-based methods, ensemble methods, and neural networks. The choice of method depends on the structure of the data, the number of observations, the degree of class imbalance, and operational requirements [3, 5, 7].

Logistic Regression is frequently used as a baseline because it is relatively simple, computationally efficient, and interpretable. Decision Trees provide explicit decision rules and can model nonlinear relationships, but individual trees may be unstable and susceptible to overfitting. KNN makes decisions based on the similarity between observations and can represent local patterns, although prediction can become computationally expensive as the dataset grows. SVM seeks a decision boundary that maximizes the margin between classes and can model nonlinear boundaries through kernel functions [7-10]. The literature also shows that class imbalance is a central issue in fraud detection. If legitimate transactions greatly outnumber fraudulent transactions, a classifier can appear accurate while failing to detect a substantial proportion of fraud cases. Techniques such as undersampling, oversampling, SMOTE, class weighting, and cost-sensitive learning have therefore been used to improve minority-class detection [6, 11, 12]. This makes recall, precision, F1-score, and ROC-AUC important complements to accuracy.

2.3 Empirical Evidence

Bhattacharyya et al. [3] found that different machine learning methods exhibit different strengths in credit card fraud classification and argued for comparative evaluation rather than assuming that one algorithm is universally superior. Carcillo et al. [13] further demonstrated the importance of scalable approaches for streaming fraud detection, highlighting the gap between experimental classification and operational fraud-monitoring environments. Research on cost-sensitive learning also indicates that the economic

consequences of false negatives and false positives should be considered when designing fraud classifiers [6].

The literature therefore suggests that fraud detection is best viewed as a multi-objective classification problem. A model that achieves the highest accuracy may not be the most useful if it misses too many fraudulent cases. Similarly, a model with high recall may generate too many false alarms. The present study extends this comparative perspective by evaluating four accessible supervised learning algorithms using several complementary metrics on the same transaction data.

III. METHODOLOGY

3.1 Research Design

The study adopted an experimental and comparative research design. The central procedure was to prepare the transaction data, train four supervised machine learning algorithms under comparable conditions, test them on unseen records, and compare their results using predefined performance measures. An Agile development approach was used in the underlying project to support iterative development, testing, evaluation, and refinement of the fraud-detection workflow.

3.2 Dataset

The study used a publicly available, anonymized credit card transaction dataset obtained through Kaggle. The dataset contains 2,000 transaction records, which were used for the development and evaluation of the proposed fraud detection models. The dataset contains transaction-related attributes that provide information about the characteristics and context of each transaction. These attributes include transaction amount, transaction channel, merchant category, device type, location state, transaction hour, international transaction status, and fraud label.

The Fraud_Label serves as the target variable for the classification task. A value of 0 represents a legitimate transaction, while a value of 1 represents a fraudulent transaction. The dataset was prepared for machine learning through appropriate preprocessing procedures, including the numerical transformation of categorical variables and feature scaling where required. Following preprocessing, the dataset was divided into 80% training data (1,600 transactions) and 20% testing data (400 transactions). The training set was used to develop the machine learning models, while the testing set was reserved for evaluating their performance on previously unseen transactions.

3.3 Data Preprocessing

Preprocessing was performed to make the transaction records suitable for machine learning. Categorical responses represented by Y and N were encoded numerically as 1 and 0. The data were then divided into training and testing portions, with 80% allocated to training and 20% to testing. The project reports 1,600 training transactions and 400 testing transactions for the described 2,000-record dataset. The split was designed to preserve a reasonable representation of the classes. Numerical variables were also scaled where required because distance- and margin-based algorithms can be influenced by differences in feature magnitude [14].

3.4 Model Development

Four supervised classifiers were selected: KNN, Decision Tree, SVM, and Logistic Regression. KNN was configured with $k = 5$. The Decision Tree used a maximum depth of 5. SVM and Logistic Regression used balanced class weights to reduce the effect of class imbalance. All models were trained using the same prepared training data and evaluated against the same testing data, allowing their results to be compared under consistent conditions [14].

3.5 Evaluation Metrics

Accuracy was used to measure the proportion of correctly classified observations. Precision measured the proportion of transactions predicted as fraudulent that were actually fraudulent. Recall measured the proportion of actual fraudulent transactions that were successfully identified. F1-score provided a combined measure of precision and recall, while ROC-AUC measured the model's ability to discriminate between the two classes over different classification thresholds. Training time was also recorded to provide an indication of computational cost.

© AUG 2026 | IRE Journals | Volume 10 Issue 2 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I2-1722685

For fraud detection, recall is particularly important because a false negative represents a fraudulent transaction that has not been identified. Nevertheless, maximizing recall without considering precision can lead to excessive false alarms. The use of multiple metrics therefore provides a more balanced basis for selecting a model [6, 12].

IV. RESULTS AND DISCUSSION

4.1 Model Performance

Table 1. Performance comparison of the evaluated machine learning models

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC	Training time (s)
KNN	0.6133	0.6279	0.6750	0.6506	0.6559	0.0067
Decision Tree	0.5400	0.6038	0.4000	0.4812	0.5473	0.0058
SVM	0.5600	0.6029	0.5125	0.5541	0.6507	0.0345
Logistic Regression	0.6333	0.6761	0.6000	0.6358	0.6855	0.0153

The results show clear differences among the four algorithms. Logistic Regression produced the highest accuracy (63.33%), precision (67.61%), and ROC-AUC (0.6855). KNN achieved the highest recall (67.50%) and F1-score (0.6506). SVM recorded moderate performance, while Decision Tree achieved the shortest training time but the weakest predictive results [14].

4.2 Accuracy and Discrimination

Logistic Regression recorded the highest accuracy at 63.33%, followed by KNN at 61.33%, SVM at 56.00%, and Decision Tree at 54.00%. The difference indicates that Logistic Regression provided the strongest overall correctness on the reported test data. However, the relatively modest accuracy values also show that none of the evaluated classifiers can be considered highly reliable for direct operational deployment without further improvement.

The ROC-AUC results reinforce the relative advantage of Logistic Regression. Its score of 0.6855 was higher than KNN (0.6559), SVM (0.6507), and Decision Tree (0.5473). The result suggests that Logistic Regression had the strongest class-discrimination ability among the evaluated models, although an AUC below 0.70 indicates substantial room for improvement.

4.3 Recall and F1-score

KNN recorded the highest recall at 67.50%. In practical terms, this means it identified a larger proportion of the fraudulent cases represented in the test data than the other evaluated models. Its F1-score of 0.6506 was also the highest, indicating the strongest balance between precision and recall among the four classifiers. Logistic Regression followed closely with an F1-score of 0.6358 and recall of 60.00% [14].

The difference between Logistic Regression and KNN illustrates why a single metric can produce a misleading conclusion. Logistic Regression was stronger on accuracy, precision, and ROC-AUC, whereas KNN was stronger on recall and F1-score. If the principal objective were to minimize missed fraud cases, KNN would deserve particular consideration. If a broader balance of correctness, precision, and discrimination is required, Logistic Regression provides the stronger overall profile.

4.4 Computational Performance

Decision Tree required the shortest reported training time at 0.0058 seconds, followed by KNN at 0.0067 seconds, Logistic Regression at 0.0153 seconds, and SVM at 0.0345 seconds. Although these differences are small for the experimental dataset, they illustrate the computational trade-off between model structure and predictive performance. Decision Tree was computationally efficient but produced substantially weaker predictive results. SVM required the longest training time while delivering only moderate classification performance [14].

4.5 Comparative Discussion

The overall results are consistent with the broader literature in one important respect: no single algorithm dominates every evaluation criterion.

Previous studies have emphasized the importance of considering the cost of classification errors and the characteristics of the underlying data [3, 6]. The present findings similarly show that Logistic Regression and KNN have different practical strengths. Logistic Regression is comparatively interpretable and achieved the best accuracy, precision, and ROC-AUC, while KNN was more sensitive to the positive class and achieved the best recall and F1-score.

The relatively weak performance of Decision Tree is also informative. Decision Trees are attractive because their decision rules are easy to interpret, but an individual tree can be unstable and may have difficulty representing complex transaction patterns without careful tuning. The 40.00% recall reported in this experiment is particularly concerning for fraud detection because it indicates that a considerable proportion of fraudulent cases were missed. The result supports the need for model tuning, class-balancing strategies, and stronger feature representations.

The findings should not be interpreted as evidence that Logistic Regression is universally superior to more advanced fraud-detection methods. Rather, it was the strongest of the four algorithms under the conditions of this experiment. The literature contains evidence that ensemble and more sophisticated learning methods can perform strongly when adequate data, feature engineering, imbalance handling, and tuning are available [6, 13]. Thus, the present result is best understood as a comparative finding within the study's selected algorithms and dataset.

V. CONCLUSION AND RECOMMENDATIONS

5.1 Conclusion

This study comparatively evaluated KNN, Decision Tree, SVM, and Logistic Regression for credit card fraud detection in an e-commerce context. The results demonstrate that model evaluation should consider several complementary metrics rather than accuracy alone. Logistic Regression achieved the highest accuracy, precision, and ROC-AUC, making it the strongest overall model among the four evaluated classifiers. KNN achieved the highest recall and F1-score, making it particularly attractive when identifying as many fraudulent transactions as possible is the primary concern.

Despite these comparative strengths, the reported performance of all four models remains moderate. The findings therefore do not support immediate deployment in a real-world financial environment without further validation. Instead, the results provide an empirical basis for selecting and improving models and demonstrate the importance of using representative transaction data and fraud-specific evaluation measures.

5.2 Recommendations

First, future studies should use larger and more representative Nigerian transaction datasets containing verified fraud labels. Second, additional transaction and behavioural features should be incorporated through systematic feature engineering. Third, class-imbalance methods such as SMOTE, carefully controlled undersampling, class weighting, or cost-sensitive learning should be investigated where appropriate [11, 12]. Fourth, hyperparameter optimization and cross-validation should be used to improve model reliability. Fifth, future research should compare the selected models with ensemble and deep-learning approaches, including Random Forest, gradient boosting, and neural networks. Finally, any operational implementation should incorporate continuous monitoring and periodic retraining because fraud patterns can change over time.

5.3 Contribution to Knowledge

The study contributes an empirical comparison of four accessible supervised learning algorithms using multiple fraud-detection metrics. It demonstrates that the best model can depend on the evaluation criterion: Logistic Regression was strongest overall, while KNN was strongest in recall and F1-score. The study also highlights the practical trade-off between predictive performance and training time and reinforces the importance of treating fraud detection as a problem in which false negatives and false positives have different consequences.

REFERENCES

[1] Adepoju, O., Wosowei, J., Lawte, S., & Jaiman, H. (2019). Comparative evaluation of credit card fraud detection using machine learning techniques. In *2019 Global Conference for Advancement in Technology (GCAT)* (pp. 1–6). IEEE. Crossref

[2] Bolton, R. J., & Hand, D. J. (2002). Statistical fraud detection: A review. *Statistical Science*, 17(3), 235–255. Crossref

[3] Bhattacharyya, S., Jha, S., Tharakunnel, K., & Westland, J. C. (2011). Data mining for credit card fraud: A comparative study. *Decision Support Systems*, 50(3), 602–613. Crossref

[4] Ngai, E. W. T., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision Support Systems*, 50(3), 559–569. Crossref

[5] Bahnsen, A. C., Aouada, D., & Ottersten, B. (2015). Example-dependent cost-sensitive decision trees. *Expert Systems with Applications*, 42(19), 6609–6619. Crossref

[6] Carcillo, F., Dal Pozzolo, A., Le Borgne, Y.-A., Caelen, O., Mazzer, Y., & Bontempi, G. (2018). SCARFF: A scalable framework for streaming credit card fraud detection with Spark. *Information Fusion*, 41, 182–194. Crossref

[7] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297. Crossref

[8] Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1), 21–27. Crossref

[9] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. Crossref

[10] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5), 1189–1232. Crossref

[11] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321–357. Crossref

[12] Dal Pozzolo, A., Caelen, O., Johnson, R. A., & Bontempi, G. (2015). Calibrating probability with undersampling for unbalanced classification. In *2015 IEEE Symposium Series on Computational Intelligence* (pp. 159–166). IEEE. Crossref

[13] Carcillo, F., Le Borgne, Y.-A., Caelen, O., Kessaci, Y., Oblé, F., & Bontempi, G. (2021). Combining unsupervised and supervised learning in credit card fraud detection. *Information Sciences*, 557, 317–331. Crossref

[14] Uchechukwu Sophia, E. (2026). *Fraud detection system using machine learning models in e-commerce* (Unpublished master's thesis). Niger Delta University.