International Journal for Electronic Crime Investigation

ISSN: 2522-3429 (Print)
ISSN: 2616-6003 (Online)

DOI: https://doi.org/10.54692/ijeci.2025.0902/254

Research Article

Vol. 9 issue 2 Jul-Dec 2025

Big Data Analytics and Machine Learning Techniques for Real-Time Credit Card Fraud Detection

Sadia Abbas Shah[1*], Rabia Javed[2], Fahima Tahir[3], Khansa Aatif[4], and Wajeeha Malik[5]

[1] School of system and technology, Department of Software Engineering, University of management and technology, Lahore, Pakistan,
[2,3,4,5] Department of Computer Science, Lahore College for Women University, Lahore, Pakistan,

Corresponding Author: sadia.abbas@umt.edu.pk

Received: Sep 15,2025; **Accepted:** Sep 27,2025; **Published:** Oct 16,2025

ABSTRACT

Big Data is commonly characterized by the 4 V's: Volume, Variety, Velocity, and Veracity. In today's digital age, data is generated in terabytes and petabytes, far exceeding the storage capabilities of a single machine. With data constantly circulating across cloud platforms, the risk of leakage and fraud has increased significantly, with credit card fraud being one of the most pressing global concerns. As numerous shopping platforms and businesses operate around us, each domain generates vast amounts of data, often reaching into yottabytes. Manually handling, analyzing, or detecting anomalies in such large-scale data is extremely challenging. However, with the advancement of computing and emerging technologies, detecting fraud has become much more efficient and scalable. This study examines the application of big data in analyzing credit card consumer behavior, specifically in the context of online transactions, password creation, age, income, and other relevant factors. The focus is on identifying anomalies in these data points to detect potentially fraudulent activities quantitative approach is employed to identify statistical patterns, and the performance of seven different machine learning algorithms, such as Logistic Regression, K-Nearest Neighbors (KNN), and XGBoost, is evaluated for their effectiveness. As technology advances, factors such as age and increasing reliance on online transactions, e-commerce, and digital banking contribute to rising vulnerabilities, making fraud detection more critical than ever. In the Real-time credit card fraud detection using big data, different algorithms are discussed and implemented so XGBOOST gives better results with 99% accuracy as another ML Algorithm. The impact of compliance on sophisticated data-based security systems will be examined in a later study, which can make use of historical fraud typologies and trends to comprehend potential shifts over time.

Keywords: Anomaly Detection, Machine Learning Algorithms, Big Data, Credit Card Fraud Detection, XGBoost and KNN

1. INTRODUCTION

Over all the world consumers now utilize the internet more frequently for banking, mobile payments, and purchases, which is convenient but also exposes them to hackers. The existing methods for identifying fraud are no longer effective; new, more effective methods that might work for the new customer type are required [1]. Conventional decision-making usually depends on empirical techniques, a specialist's knowledge, or strict, antiquated algorithms that are unable to adjust to complex patterns or new information in contemporary data. Even while the aforementioned techniques might be effective in some fields, they are frequently limited by human cognitive biases and the sheer difficulty or impossibility of processing such massive amounts of incoming data. Delays in response, significantly increased error rates, and a general inability to handle massive datasets to reveal accurate and important information are all possible outcomes of decisions made using traditional methods rather than machine learning techniques [2]. Consumers benefit from credit cards in addition to debit cards since they protect items that may be lost, stolen, or destroyed. Before using their credit card to make any purchases, customers must confirm the transaction with the retailer [3]. Credit card fraud is detected using a variety of techniques, such as statistical, machine learning, and deep learning methods. To find and examine irregularities in credit card transactions, statistical methods, including regression, hypothesis testing, and clustering are used. Machine learning techniques, on the other hand, use algorithms to analyze previous data and identify fraudulent activities in real time. Neural networks are used in deep learning techniques to automatically find complex patterns and features in large, complicated datasets, leading to incredibly accurate fraud detection [4]. We examine the effectiveness of artificial neural networks (ANN-DL), K-nearest neighbors (KNN), Naïve Bayes classifiers, decision trees, random forest classifiers, and logistic regression. To find the best models for identifying fraudulent transactions, we compare important performance parameters like accuracy, sensitivity, specificity, and F1-score. Furthermore, we investigate how various folds in cross-validation affect model performance, offering information on the classifiers' stability and resilience. This study adds to the continuous efforts to create reliable and effective fraud detection systems, providing insightful information to researchers and financial organizations working to successfully battle credit card theft [5].

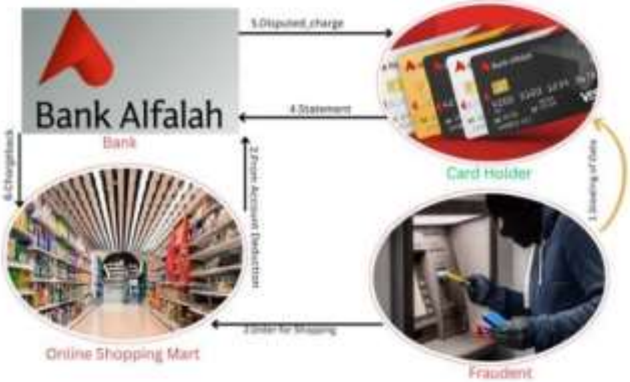


Figure 1. Large-scale real-time credit card fraud

Fig 1 illustrates the process involving a transaction dispute related to fraudulent activity using a credit card issued by Bank Alfalah. Initially, a cardholder engages in an online shopping order, which leads to a deduction from their account. However, the cardholder later receives a statement indicating a disputed charge, prompting them to seek a chargeback from the bank. The flow also highlights the involvement of a fraudulent entity that compromises the cardholder's data, leading to unauthorized transactions. This visualization encapsulates the cycle of online shopping, potential fraud, and the subsequent actions taken by the cardholder and bank to resolve the dispute.

2. RELATED WORK

Machine learning solutions are extremely helpful in various effective spheres in which data have to be processed; one of them is the detection of card fraud. Some of the methods recommended in prior studies have proposed inclusion of methods to detect fraud during the supervised approaches, the unsupervised approaches, and even a hybrid approach; this make sit necessary and important to know some technology in the identification of credit card fraud and understand better the nature of card frauds. Numerous measures were proposed and verified. The following brief will review most of them. Prediction of card fraud has been centered on the interpretation of the card actions during purchase. During the process of identifying card fraud, most of them were put in place, including neural network (NN), genetic algorithm (GA), support vector machine (SVM), frequent itemset mining (FISM), decision tree (DT), optimization algorithm of the migratory birds (MBO) and naive Bayes process (NB). In its performance, the quantitative analysis carried out is an estimation of the logistic regression and naive Bayes analysis alongside assessment of the Bayesian and neural system output on the dataset of credit card fraud [6]. A synopsis of the available works done towards real-time credit card fraud detection is contained in the Table 1.

Kasongo [12] designed a GA-based FS to enhance the performance of ML-based models used in the field of intrusion detection systems. The results demonstrated that the RF classifier performed better when GA was used, with an Area Under the Curve (AUC) of 0.98.

Numerous assessments of earlier studies have been carried out to analyze machine learning applications in detecting fraudulent credit card activity using innovative and stacked architectures [13–16]. Numerous studies have been completed using various data mining approaches; according to reviews [17–21], over 23% of these studies have focused on the SVM methodology, with the naïve Bayes and random forest techniques accounting for 13% of the total number of research papers. Up until now, the research has mostly concentrated on the data. The researchers found that the data was unbalanced, which is likely what caused the models' performance to deteriorate. As a result, they used under-sampling and oversampling, and the under-sampling strategy using logistic regression produced better results [22]. Researchers have successfully experimented with artificial neural networks for fraud detection since they perform well when dealing with complex data [23]. In this paper, additional analysis has been demonstrated to examine more commonly used techniques that may perform better than the earlier findings. After applying 19 resampling techniques to each algorithm, the top three are chosen to be used in the second stag

Table 1. The recent efforts to improve the identification of real-time credit Card Fraud

Reference	Dataset	Method	Pros	Cons	Accuracy
[7]	European card holders	KNN	Mean-squared error is decreased using CFLANN.	KNN takes a lot of time.	97.56% for identifying fraudulent transactions
[8]	Chinese financial institution	Support vector machine, neural network	The CCFD performs better overall.	lengthy procedure	95.20% recall and 99.21% accuracy
[9]	Datasets from financial institutions	SVM, KNN	It facilitates the categorization of real-world interactions.	The algorithm needs in-depth knowledge to make predictions in real-world scenarios.	SVM's 91% accuracy and KNN's 72%
[10]	Banks datasets	SVM	SVM avoids overfitting.	Model training takes a lot of time.	When compared to the hybrid B.P. model, SVM performs well.
[11]	UCSD FICO datasets	VM, KNN, Naive Bayes (NB)	The less significant change has little effect on how the model is implemented.	Model training takes a lot of time. The prediction is not always correct. KNN is susceptible to dataset noise.	SVM achieved 20% accuracy, NB 15%, and KNN 10%.

Our objective is to address three main problems with credit card fraud datasets: substantial class imbalance, inclusion of labelled and unlabeled samples, and increased processing capacity. A variety of supervised and semi-supervised machine learning techniques are used for fraud detection.

3. Materials and Methods

Figure 2 illustrates the proposed workflow for real-time credit card fraud detection is structured in several key stages. The Input Layer involves gathering data from various transaction features, including timestamps and fraud indicators. Following this, the Preprocessing phase applies the feature

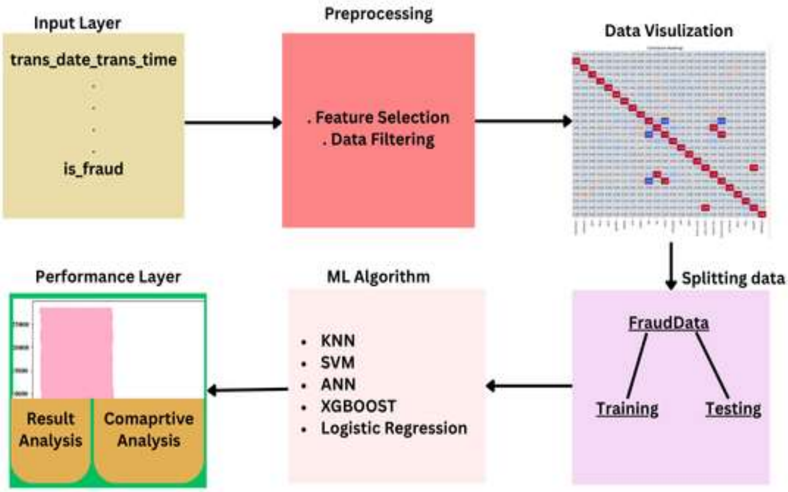
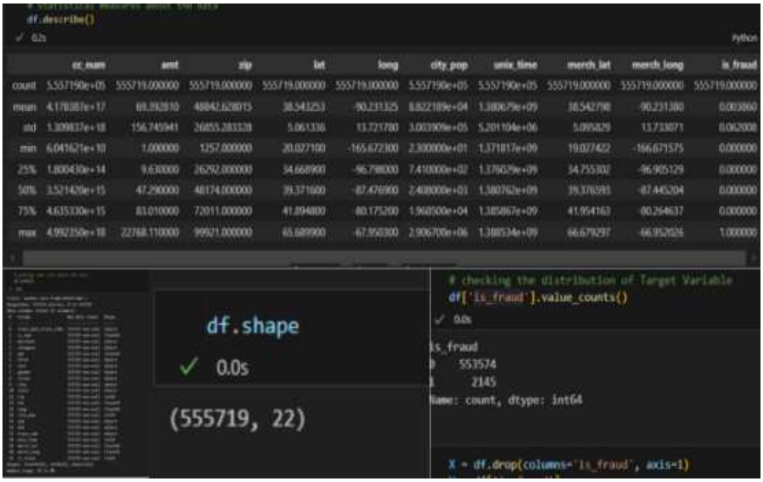


Figure 2. Workflow of Real-Time Credit Card Fraud Detection

Selection and data filtering techniques to refine the dataset, enhancing its quality for analysis. This groundwork is followed by Data Visualization, where the relationships among features are explored using visualization tools to identify patterns and insights. Subsequently, the ML Algorithm step employs various machine learning techniques such as KNN, SVM, ANN, XGBoost, and Logistic Regression to develop models that can predict fraudulent activities. Finally, in the Performance Layer, both result analysis and comparative analysis are conducted to evaluate the effectiveness of the models, as the dataset is split into training and testing subsets for robust performance assessment. **Input Layer** shows the Dataset taken from Kaggle [24] with 22 columns, in which the target column is the 'is fraud' attribute from the Kaggle file with the name of 'fraudTest.csv'. This dataset of simulated credit card transactions includes both authentic and fraudulent transactions. **Data Preprocessing** stage has a big impact on how machine learning models are used later. Some negative data features, like as noise, excessive dimensionality, and outliers, can negatively affect model performance, and many models are unable to handle missing values. Therefore, the dataset is improved in terms of accuracy and completeness by doing data preparation. Figure 3 shows the data handling, filtering, and managing data, calculating statistical values of fraud data implemented with Python on Visual Studio Code. Python is installed on an HP EliteBook equipped with an 11th-generation Intel Core i7 processor and 8 GPU cores



	cc_num	amt	zip	lat	long	city_pop	unix_time	merch_lat	merch_long	is_fraud
count	5.557190e+05	555719.000000	555719.000000	555719.000000	555719.000000	5.557190e+05	5.557190e+05	555719.000000	555719.000000	555719.000000
mean	4.178387e+17	69.392810	48842.628015	38.543253	-90.231325	8.822189e+04	1.380679e+09	38.542798	-90.231380	0.003860
std	1.309837e+18	156.745941	26855.283328	5.061336	13.721780	3.003909e+05	5.201104e+06	5.095829	13.733071	0.062008
min	6.041621e+10	1.000000	1257.000000	20.027100	-165.672300	2.300000e+01	1.371817e+09	19.027422	-166.671575	0.000000
25%	1.800430e+14	9.630000	26292.000000	34.668900	-96.798000	7.410000e+02	1.376029e+09	34.755302	-96.905129	0.000000
50%	3.521420e+15	47.290000	48174.000000	39.371600	-87.476900	2.408000e+03	1.380762e+09	39.376593	-87.445204	0.000000
75%	4.635330e+15	83.010000	72011.000000	41.894800	-80.175200	1.968500e+04	1.385867e+09	41.954163	-80.264637	0.000000
max	4.992350e+18	22768.110000	99921.000000	65.689900	-67.950300	2.906700e+06	1.388534e+09	66.679297	-66.952026	1.000000

Figure 3. Exploratory Data Analysis for Real-Time Debit Card Fraud Detection

Data Visualization works as a Exploratory Data Analysis (EDA) which is performed on the Fraud Test dataset. Relevant libraries are imported sequentially for data loading, visualization, and normalization. To gain initial insights, we examine the figure 3 shows dataset's column names, shape, and statistical summaries of each feature. Figure 4 shows data distribution as his plot, pair plot and relational plot.

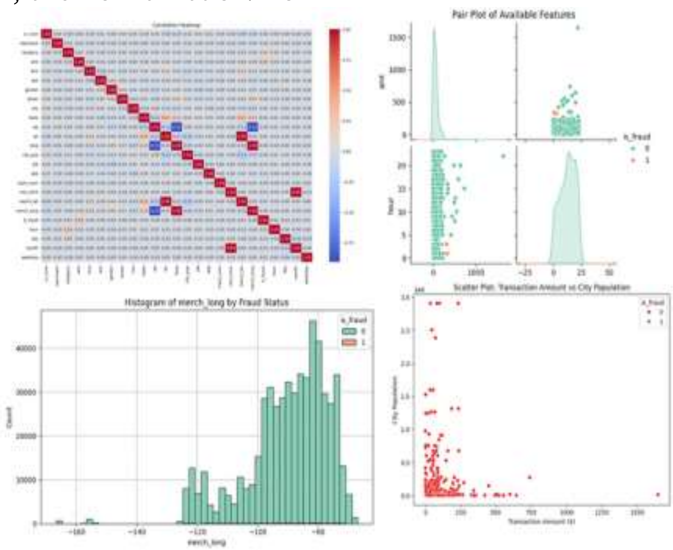


Figure 4. Data Distribution for Real-Time Credit Card Fraud Detection.

4. Results and experiments

The dataset's balanced distribution is depicted in Figure 5. It has been the most crucial issue to address to identify fraudulent behavior. Since there are very few fraudulent situations, any algorithm may assume that any transaction requested from the database will be typical. But in practice, it isn't the case.

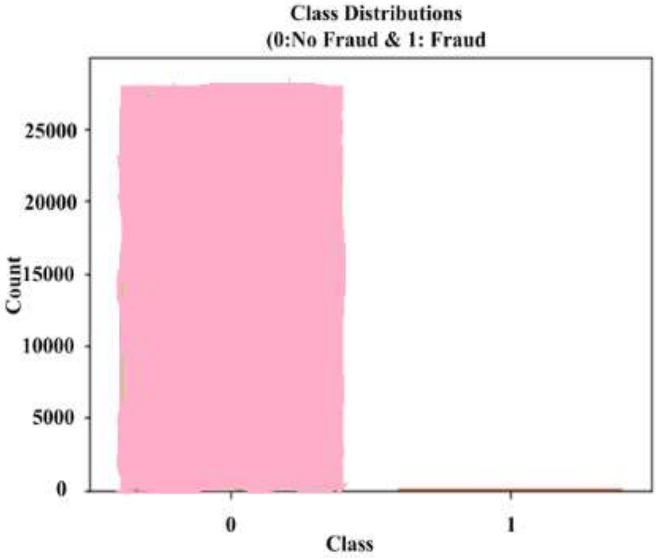


Figure 5. An unbalanced dataset

Table 2. Performance of 7 Techniques for Real Credit Card Fraud Detection

Techniques	**Accuracy**	**Precision**	**MCC**
KNN	0.97	0.97	0.96
SVM	0.97	0.77	0.53
Logistic Regression	0.98	0.88	0.68
XGBoost	0.99	0.98	0.97
ANN	0.93	0.97	0.86
Randomforest	0.96	0.89	0.83
DecisionTree	0.96	0.94	0.84

In the provided Table 2, the performance of seven techniques for real credit card fraud detection is evaluated based on three metrics: Accuracy, Precision, and MCC (Matthews Correlation Coefficient). The KNN (K-Nearest Neighbors) model demonstrates the highest overall effectiveness with an accuracy of 0.99, precision of 0.97, and MCC of 0.96, indicating its reliability in correctly identifying fraudulent transactions. Following closely is the SVM (Support Vector Machine) with an accuracy of 0.9972, though it shows a relatively lower precision of 0.77 and MCC of 0.53, suggesting issues with false positives. Logistic Regression and XGBoost exhibit comparable performance, with Logistic Regression achieving 0.991 accuracy and 0.88 precision, while XGBoost scores an accuracy of 0.99 and precision of 0.98. The ANN (Artificial Neural Networks) technique yields slightly lower results with an accuracy of 0.93 and a precision of 0.97. The Random Forest model shows an accuracy of 0.96 and a precision of 0.89, while the Decision Tree model has the same accuracy of 0.96 but a higher precision at 0.94. This comparison highlights varying levels of effectiveness among the techniques, with KNN and XGBoost emerging as the strongest contenders for accurately detecting fraudulent activities in credit card transactions.

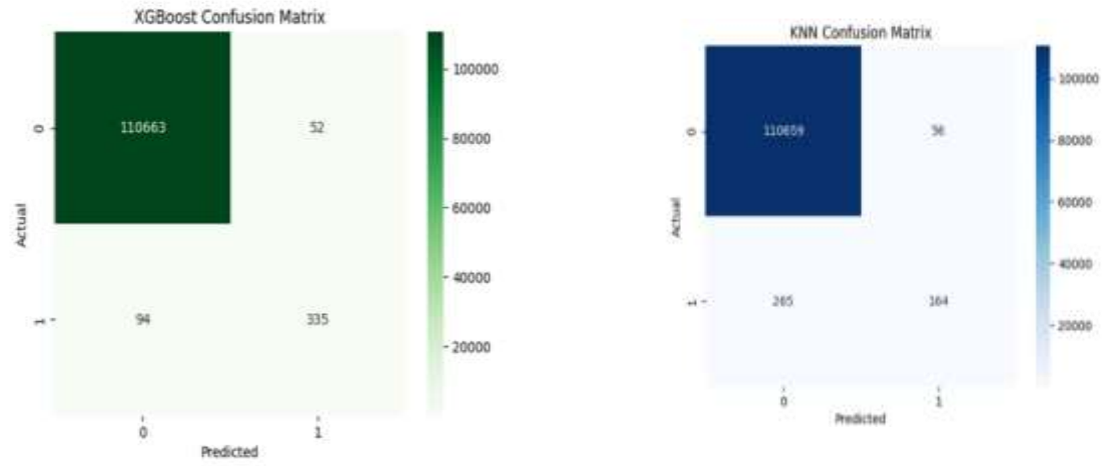


Figure 6. Confusion Matrix for Real-Time Credit Card Fraud Detection

Figure 6. displays comparison confusion matrices for two machine learning models: XGBoost and K-Nearest Neighbors (KNN). Each matrix quantifies the performance of its respective model by categorizing actual and predicted classifications into four distinct areas: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Comparatively, both models exhibit high true positive rates, indicating that they effectively

identify positive cases. However, XGBoost demonstrates slightly better performance, with fewer false positives and false negatives than KNN. This suggests that while both models are competent, the XGBoost model may provide more accurate predictions in this context.

$$PRECISION = \frac{TP}{TP + FP} \tag{1}$$

$$ACCURACY = \frac{TN+TP}{TN+TP+FN+FP} \tag{2}$$

$$MCC = \frac{TP*TN-FP*FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{3}$$

5. CONCLUSION

Frequent occurrences of credit card fraud led to significant financial losses. Online credit card transactions account for a significant portion of the vastly increasing number of transactions that take place online. As a result, banks and other financial organizations provide credit card fraud detection software that is highly valued and in high demand. Transactions that are fraudulent can take many different forms and fall under a variety of headings. This essay focuses on four primary instances of fraud in actual transactions. A number of machine learning models are used to tackle each scam, and an evaluation is used to determine which approach works best. This assessment offers a thorough manual for choosing the best algorithm based on the kind of frauds, and we use a suitable performance metric to demonstrate the assessment. Another issue that concerns us in our research is real time credit card fraud detection. With such a question as to whether a particular transaction is genuine or is fraudulent, we resort to predictive analytics that is undertaken by the machine learning models that have been implemented along with an API module. We assess as well a new method which effectively addresses the skewed distribution of the data. A financial institution supplied us with the data used in our experiments under a confidential disclosure accord. KNN and XGBoost are the two most suitable to detect the fraudulentness in real-time, specifically in credit cards, where there is some fraud conduct.

6. REFERENCES

[1] F. T. Zohora, R. Parveen, A. Nishan, M. R. Haque, and S. Rahman, “Optimizing credit card security using consumer behavior data: A big data and machine learning approach to fraud detection,” *Frontline Marketing Management Economics Journal*, vol. 4, no. 12, pp. 26–60, 2024.

[2] L. Theodorakopoulos, A. Theodoropoulou, F. Zakka, and C. Halkiopoulos, “Credit card fraud detection with machine learning and big data analytics: A PySpark framework implementation,” 2024.

[3] Y. Zou and D. Cheng, “Effective high-order graph representation learning for credit card fraud detection,” *arXiv preprint* arXiv:2503.01556, 2025.

[4] A. R. Khalid *et al.*, “Enhancing credit card fraud detection: an ensemble machine learning approach,” *Big Data and Cognitive Computing*, vol. 8, no. 1, p. 6, 2024.

[5] S. F. Farabi *et al.*, “Enhancing credit card fraud detection: A comprehensive study of machine learning algorithms and performance evaluation,” *Journal of Business and Management Studies*, vol. 6, no. 3, pp. 252–259, 2024.

[6] M. Yousif *et al.*, “Machine learning-based suicide risk assessment and intervention strategies for depression,” 2024.

[7] Y. Wang, “A data balancing and ensemble learning approach for credit card fraud

detection," in *Proc. 4th Int. Symp. Computer Applications and Information Technology (ISCAIT)*, IEEE, Mar. 2025, pp. 386–390.
[8] L. Bonde and A. K. Bichanga, "Improving credit card fraud detection with ensemble deep learning-based models: A hybrid approach using SMOTE-ENN," *J. Comput. Theor. Appl.*, vol. 2, no. 3, pp. 383–394, 2025.
[9] E. F. Aminu *et al.*, "A hybridized SMOTE-ENN approach on imbalanced dataset of fraudulent credit-card scenario," *i-Manager's Journal on Data Science & Big Data Analytics (JDS)*, vol. 3, no. 1, 2025.
[10] M. Zhu, Y. Zhang, Y. Gong, C. Xu, and Y. Xiang, "Enhancing credit card fraud detection: A neural network and SMOTE integrated approach," *arXiv preprint* arXiv:2405.00026, 2024.
[11] C. Yu *et al.*, "Credit card fraud detection using advanced transformer model," in *Proc. IEEE Int. Conf. Metaverse Computing, Networking, and Applications (MetaCom)*, Aug. 2024, pp. 343–350.
[12] P. Chatterjee, D. Das, and D. B. Rawat, "Digital twin for credit card fraud detection: Opportunities, challenges, and fraud detection advancements," *Future Generation Computer Systems*, vol. 158, pp. 410–426, 2024.
[13] K. G. Dastidar, O. Caelen, and M. Granitzer, "Machine learning methods for credit card fraud detection: A survey," *IEEE Access*, 2024.
[14] F. D. Data, C. S. Subrahmanyam, N. Deshai, and K. S. J. T. Rajesh, "Algorithms in advanced artificial intelligence," in *Algorithms in Advanced Artificial Intelligence: ICAAAI-2023*, R. N. V. Jagan Mohan et al., Eds., London, U.K.: Taylor & Francis Group, 2024, p. 456.
[15] F. D. Data, C. S. Subrahmanyam, N. Deshai, and K. S. J. T. Rajesh, "Algorithms in advanced artificial intelligence," in *Algorithms in Advanced Artificial Intelligence: ICAAAI-2023*, R. N. V. Jagan Mohan et al., Eds., London, U.K.: Taylor & Francis Group, 2024, p. 456.
[16] J. Liu, X. Zhang, and H. Xiong, "Credit risk prediction based on causal machine learning: Bayesian network learning, default inference, and interpretation," *Journal of Forecasting*, vol. 43, no. 5, pp. 1625–1660, 2024.
[17] Z. Wang, Q. Shen, S. Bi, and C. Fu, "AI empowers data mining models for financial fraud detection and prevention systems," *Procedia Computer Science*, vol. 243, pp. 891–899, 2024.
[18] H. Chen *et al.*, "The role of blockchain in finance beyond cryptocurrency: trust, data management, and automation," *IEEE Access*, vol. 12, pp. 64861–64885, 2024.
[19] I. Y. Hafez *et al.*, "A systematic review of AI-enhanced techniques in credit card fraud detection," *Journal of Big Data*, vol. 12, no. 1, p. 6, 2025.
[20] T. A. Gaav, H. U. Adoga, and T. Moses, "Recent advances in credit card fraud detection: An analytical review of frameworks, methodologies, datasets, and challenges," *J. Future Artif. Intell. Technol.*, vol. 2, no. 3, pp. 343–369, 2025.
[21] F. Moradi, M. Tarif, and M. Homaei, "A systematic review of machine learning in credit card fraud detection," *MDPI Preprints*, 2025.
[22] L. Theodorakopoulos, A. Theodoropoulou, A. Tsimakis, and C. Halkiopoulos, "Big data-driven distributed machine learning for scalable credit card fraud detection using PySpark, XGBoost, and CatBoost," *Electronics*, vol. 14, no. 9, p. 1754, 2025.
[23] J. Wang, J. Liu, W. Zheng, and Y. Ge, "Temporal heterogeneous graph contrastive learning for fraud detection in credit card transactions," *IEEE Access*, 2025
[24] K. Shenoy, "Credit card transactions fraud detection dataset," *Kaggle*, 2020.