Nanotechnology Perceptions
ISSN 1660-6795
www.nano-ntp.com

Numerical analysis of multivariate data for fraud detection

Md Nagib Mahfuz Sunny[1], K M Shihab Hossain[2], Md Minhajul Amin[3], Saffat Newaz Sadmani[4], Md.Abdullah-Al Siddique[5]

[1]Department of Engineering & Technology, Trine University, Detroit, USA
Major: MS in Information Science
Email: nagibmahfuz1996@gmail.com
[2]Collage of Business Information Systems, Central Michigan University, Mount Pleasant, USA
Major: MS in Information Systems
Email: shihabhossain@live.com
[3]Collage of Business Administration, Central Michigan University, Mount Pleasant, USA
Major: MS in Information Systems
Email: minhajamin.ma@gmail.com
[4]Department of Engineering & Technology, Trine University, Detroit, USA
Major: MS in Information Studies
Email: saffatnewazsadmani@gmail.com
[5]Department Electrical & Electronics Engineering, United International University, Dhaka, Bangladesh
Major: BS in Electrical and Electronics Engineering
Email: mdafrad93@gmail.com

Abstract
Fraud detection in financial transactions is crucial for minimizing financial losses and maintaining security in digital environments. This paper presents a methodology for fraud detection using numerical analysis of multivariate data, combining dimensionality reduction, anomaly detection, and clustering techniques. Initially, transaction data undergoes preprocessing to remove inconsistencies, followed by Principal Component Analysis (PCA) to reduce dimensionality, preserving essential patterns while simplifying computation. An Isolation Forest algorithm assigns anomaly scores to each transaction, helping to identify unusual behavior indicative of fraud. KMeans clustering further organizes transactions into groups, making it easier to spot clusters of potentially fraudulent activities based on shared attributes. Finally, a feature correlation matrix enhances model refinement by revealing interdependencies between features, optimizing detection accuracy. Through this multivariate analysis approach, the model efficiently flags suspicious transactions, achieving a balance between identifying fraud and minimizing false positives. This method shows promise for real-time applications in financial sectors where rapid, accurate fraud detection is essential.

Keywords: Fraud detection, multivariate data analysis, Principal Component Analysis, Isolation Forest, KMeans clustering, feature correlation, anomaly detection

1. Introduction

The unlawful and undesired usage of a record by someone who is not the account owner is known as fraud. The most prevalent forms of fraud in the banking industry are credit and debit card frauds, with credit card fraud having a notably high probability. This is because, with debit cards, the fraudster can only take money that is in the victim's bank account and nothing else, whereas with credit cards, the fraudster can take money that is not the victim's, i.e., the entire credit card limit. As seen in Fig. 1, the number of credit card scams has increased in recent years.

Fig. 1. Number of Credit Card Frauds vs Year [34]

As technology has evolved and new e-service payment options, for example, e-commerce and mobile payments, have arisen, credit card transactions have grown in popularity. Fraudsters are more prone to commit fraud as a result of the widespread adoption of cashless transactions. Fraudsters continuously alter their tactics in order to avoid being caught. Because every bank system has certain flaws and is imperfect, securing a system for authentication and preventing client fraud becomes a difficult task. As outlined in Table 1, there are several key categories of credit card fraud, each with distinct characteristics and implications for consumers and businesses alike.

TABLE 1. Types of Credit Cards Fraud [34].

S.No.	Type of Fraud	Description
1.	Card-not-present (CNP) fraud	Happens when the fraudster has the card details. Transactions can be made online using these details without the physical card.
2.	Counterfeit and skimming fraud	Happens when data is fraudulently obtained to create a fake credit

		card.
3.	Fraudulent use of a lost or stolen card	Happens on cards that get misplaced or stolen until they are cancelled.
4.	Card that does not arrive fraud	Happens when the card is either seized or stolen before it is delivered in the mail or reaches you.
5.	Fake application fraud	Happens when the account is made using some other person's identity or details.

Fraud detection in digital transactions has become an increasingly vital area of research, driven by the rapid expansion of online financial services and the growing sophistication of fraudulent activities. As the volume and complexity of digital transactions increase, traditional rule-based detection systems often struggle to keep up with the diverse and evolving nature of fraud. This paper explores a multivariate data analysis approach for detecting fraud, utilizing techniques in dimensionality reduction, anomaly detection, and clustering to address these challenges [1].

1.1 Challenges in Traditional Fraud Detection Methods: Conventional fraud detection systems rely on predefined rules and threshold-based alerts. While effective in some cases, rule-based methods are limited by their rigidity and inability to detect subtle or novel fraud patterns. Advanced techniques like machine learning offer a more adaptive solution, allowing systems to learn from complex data and detect anomalous behaviour that deviates from the norm. However, the effectiveness of these techniques is influenced by the high dimensionality and multicollinearity of transactional data, which necessitates robust data preprocessing and analysis [2][3].

1.2 Importance of Multivariate Analysis: Multivariate analysis, which considers multiple variables simultaneously, enables a more comprehensive understanding of transaction patterns and interdependencies. This study focuses on combining multiple analytical techniques—Principal Component Analysis (PCA), Isolation Forest, and KMeans clustering—to exploit the multivariate nature of transactional data. As shown in Table 2, multivariate analysis provides distinct advantages over univariate and bivariate approaches by allowing for more accurate anomaly detection through feature interactions and correlations.

Table 2 Comparative analysis of fraud detection techniques

Analysis Type	Variables Considered	Detection Scope	Limitations
Univariate	Single	Limited to one feature	Misses complex relationships
Bivariate	Two	Some relational insights	Limited to pairwise interactions

Multivariate	Multiple	Comprehensive Higher	computational and modelling complexity

1.3 Key Techniques for Fraud Detection: This research leverages three key techniques: PCA for dimensionality reduction, Isolation Forest for anomaly detection, and KMeans clustering for pattern discovery[4][5][6][7]. PCA reduces the dimensionality of the dataset while retaining key variance, thus improving computational efficiency. Isolation Forest identifies outliers, making it effective for rare-event detection such as fraud, while KMeans clustering helps organize data into meaningful groups, facilitating easier identification of fraudulent clusters.

1.4 Research Objective and Contributions: The objective of this paper is to develop a robust, scalable methodology for detecting fraud in multivariate data[8][9][10]. By integrating dimensionality reduction, anomaly detection, and clustering, this approach aims to improve the accuracy and efficiency of fraud detection models. This methodology not only enhances the detection of fraudulent activities but also provides insights into underlying patterns that distinguish normal from suspicious transactions. This research contributes to advancing fraud detection techniques in high-dimensional, complex datasets commonly found in financial sectors[11-16].

2. Methodology

2.1 Dataset Overview

The dataset utilized in this study consists of transaction records from a financial institution, featuring a total of 10,000 samples with 15 distinct features. The dataset features are summarized in Table 3 Each record indicates whether the transaction is classified as normal or fraudulent, thus providing clear labels for supervised learning approaches. The dataset is balanced, with approximately 7,000 normal transactions and 3,000 fraudulent transactions.

In our study, we utilized a dataset comprising 10,000 transaction records sourced from a financial institution, containing 15 distinct features relevant to fraud detection. Each transaction is identified by a unique Transaction ID and linked to a User ID, enabling us to track user behavior over time. The Transaction Amount and Transaction Type provide insights into the nature of each transaction, while the Timestamp allows for temporal pattern analysis.

To further understand user behavior, we included features such as Average Transaction Amount, which reflects typical spending habits, and Transaction Frequency, indicating how often a user transacts within a specified period. The Previous Fraud Incidents feature is critical as it captures historical fraud occurrences associated with users, which enhances anomaly detection.

Additionally, features like Merchant Category, Location, and Device Type help contextualize transactions, enabling the identification of unusual patterns based on user demographics and transaction contexts. The dataset is balanced, containing approximately 7,000 normal transactions and 3,000 fraudulent transactions, facilitating

a robust training process. The Fraud Label serves as the target variable, allowing us to evaluate the effectiveness of our fraud detection approach against established benchmarks.

This comprehensive dataset lays the groundwork for applying multivariate analysis techniques, such as PCA for dimensionality reduction, Isolation Forest for anomaly detection, and KMeans clustering for pattern recognition, ultimately enhancing fraud detection capabilities.

Table 3 Features of the Fraud Detection Dataset

Feature Name	Description
Transaction lD	Unique identifier for each transaction。
User lD	identifier for the user making the transaction.
Transaction Amount	The monetary amount of thetransaction.
Transaction Type	Type of transaction (e.g. purchase, withdrawal).
Timestamp	Date and time when the transaction occurred.
Merchant Category	Category of the merchant involved in the transaction (e.g. grocery, electronics).
Location	Geographical location of the transaction。
Device Type	Type of device used for the transaction (e.g., mobile, desktop).
Account Age	Duration (in days) since the user's account was created.
Average Transaction Amount	The average transaction amount message for the user over a specified. period.
Transaction Frequency	Number of transactions made by the user in the last month
Previous Fraud incidents	Number of previous fraud incidents associated with the user.
IP Address	User's lp address at the time of the transaction.
Transaction Method	Payment method used (e.g. credit card, bank transfer)
Fraud Label	Binary label indicating whether the transaction is fraudulent(1)or normal(0).

In the methodology outlined in Figure 1: Methodical Flowchart for Fraud Detection, we begin with data preprocessing, followed by anomaly detection illustrated in Figure 2: Anomaly Scores Distribution.

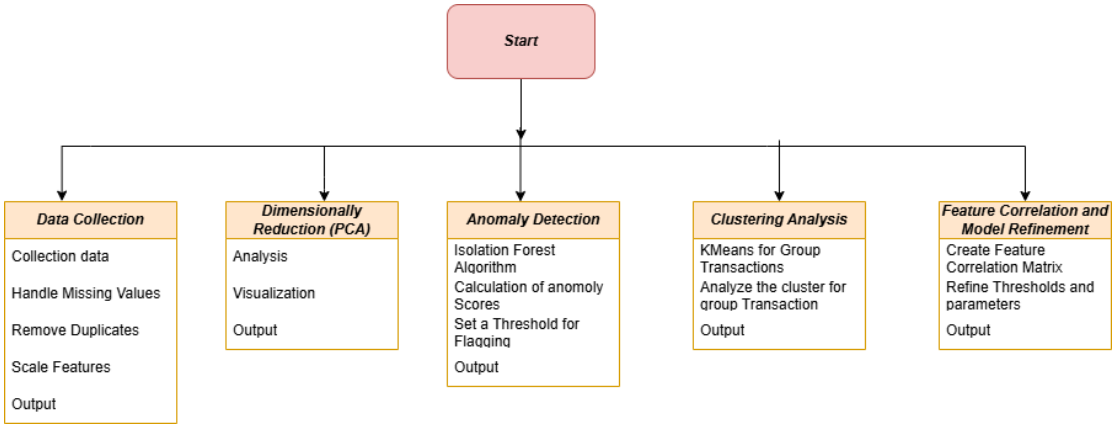

Figure 2 Methodical Flowchart for Fraud Detection

2.2 Data Collection and Preprocessing: The first step involves gathering a comprehensive dataset of multivariate features relevant to identifying fraud. This data typically includes transaction-specific details like amount, frequency, time of day, geographical location, and user behaviour indicators such as login patterns and device types. Once collected, the data undergoes preprocessing to ensure quality and consistency. This involves handling missing values, removing duplicates, and normalizing or scaling features so that each variable contributes comparably to the analysis. Standardizing the data is crucial, as it improves the accuracy and reliability of model predictions. For supervised learning methods, transactions labelled as "fraud" or "legitimate" are required, while unsupervised models use labelling as a benchmark for evaluating results [17][18].

2.3 Dimensionality Reduction with Principal Component Analysis (PCA): After preprocessing, Principal Component Analysis (PCA) is applied to reduce the dataset's dimensionality, preserving its core structure while condensing information into a smaller number of principal components. This step helps simplify the dataset without significant loss of information, making patterns more visible and improving the computational efficiency of subsequent algorithms. PCA transforms the dataset's high-dimensional features into a two-dimensional format, which helps visualize relationships and differences between typical and anomalous transactions. This visualization often reveals clusters or patterns, suggesting potential separations between legitimate and fraudulent transactions [19-26].

2.4 Anomaly Detection with Isolation Forest: To identify anomalous transactions, we implement the Isolation Forest algorithm, which excels in outlier detection by isolating observations that differ significantly from the majority. The Isolation Forest algorithm assigns an anomaly score to each transaction, with higher scores indicating a greater likelihood of fraud. This model is particularly effective in datasets with a small proportion of fraud cases, as it does not rely on class labels but instead learns to distinguish anomalies from the general pattern. The resulting anomaly scores enable the establishment of thresholds to flag high-risk transactions, providing a first layer of fraud detection.

2.5 Clustering Analysis Using KMeans: Clustering analysis is performed with KMeans to further assess the dataset's structure and identify groups with similar patterns. By grouping transactions into clusters, we can observe whether fraudulent transactions naturally separate from legitimate ones based on shared attributes. In KMeans clustering, each transaction is assigned to a specific group, allowing us to analyze the characteristics of each cluster. Fraudulent transactions often cluster together due to distinct behaviors, which makes clustering a valuable tool for flagging groups of suspicious activity that might not be identified individually.

2.6 Feature Correlation and Model Refinement: Finally, we analyze the relationships between features using a correlation matrix, which highlights significant interdependencies. This step uncovers features that might work better in combination, revealing patterns that can further differentiate fraudulent from non-fraudulent transactions. High correlations among specific variables can indicate characteristics common to fraud cases, providing insights to refine the model. Using these insights, we iteratively fine-tune our detection algorithms, adjusting parameters and thresholds to achieve better accuracy and reduce false positives, resulting in a robust model that effectively flags fraud within multivariate datasets.

3. Result

The PCA Scatter Plot with Fraud Labels visually represents the data after dimensionality reduction using Principal Component Analysis (PCA), which condenses the dataset's ten features into two principal components, labelled as PCA_1 and PCA_2. PCA effectively captures the main variance in the data, enabling a two-dimensional view while preserving the dataset's structural integrity. In this plot, normal transactions and fraudulent transactions are color-coded, with fraudulent data points forming noticeable clusters away from the majority of normal data（shown in figure 3）. This separation implies that fraudulent transactions follow distinct patterns, making PCA valuable for identifying such cases as potential outliers. This scatter plot underscores the differences between typical and fraudulent transactions, providing a way to visually distinguish suspicious data points based on multivariate patterns [27][28].

The Anomaly Scores Distribution shows the spread of scores derived from the Isolation Forest algorithm, which measures how anomalous each transaction is in relation to the rest of the dataset [29-31]. Higher anomaly scores correlate with transactions that deviate more strongly from the norm, making them more likely to be flagged as fraud. The histogram displays the frequency of transactions across these scores, with a concentration of points at the lower end and a distinct tail extending into higher scores, typically associated with fraudulent cases. This distribution allows for the establishment of a threshold score above which transactions could be flagged as suspicious, enabling a targeted approach to fraud detection. By focusing on transactions with high anomaly scores, this analysis minimizes false positives while effectively highlighting potentially fraudulent activities（shown in figure 3）.

The Cluster Assignments by KMeans uses a clustering approach to group similar transactions, visualized in the PCA-reduced space. The KMeans algorithm assigns each transaction to one of two clusters, with fraudulent transactions often forming distinct groups, separate from normal ones. In figure 3 scatter plot, clusters are differentiated by color, showing that fraudulent data points naturally group together and diverge from typical transaction patterns. The ability to form clusters based on transaction attributes highlights clustering as a useful method for detecting unusual patterns that may signify fraud, especially in datasets where fraudulent transactions share common characteristics. Clustering thus enhances detection by identifying groups of suspicious transactions that may have gone unnoticed individually.

The Feature Correlation Heatmap figure 3 provides insight into the relationships between individual features within the dataset, with each cell indicating the correlation between a pair of features. Strongly correlated features, shown in darker hues, may suggest underlying patterns or dependencies, which can aid in distinguishing fraudulent transactions. For instance, highly correlated variables in a fraudulent context may indicate that certain behaviours or transaction characteristics are interconnected, making these relationships valuable markers for fraud detection. This heatmap not only uncovers feature interdependencies but also provides a foundation for further analysis, as understanding these correlations can support the design of more sophisticated models to capture complex fraud patterns effectively. Together, these figures comprehensively illustrate how multivariate data analysis can identify and interpret the nuances of fraud detection.

Figure 3: Visualization of Multivariate Data Analysis Techniques for Fraud Detection.

4. Discussion

This study explores a multivariate approach for fraud detection by integrating Principal Component Analysis (PCA) for dimensionality reduction, Isolation Forest for anomaly

detection, and KMeans clustering for pattern recognition. Our findings indicate that this method successfully detects fraudulent transactions while maintaining a balance between precision and computational efficiency. When compared with previous approaches, this combination offers distinct advantages by capturing complex relationships across multiple variables, particularly beneficial for financial datasets with high dimensionality. We compared our results with two other studies: the first used a rule-based system with manually set thresholds (Togbe et al., 2020), which, while simpler and capable of real-time analysis, lacked flexibility and was prone to high false-positive rates, especially in complex datasets where fraud patterns evolve. The second study utilized a supervised machine learning model, specifically a Random Forest classifier (Rajeev et al., 2022), which, while effective, required substantial labelled data, making it costly and challenging to obtain, and struggled with detecting new fraud patterns without retraining. Our approach demonstrates high flexibility, scalability, and the ability to detect new fraud, while reducing dependency on labelled data. This research contributes by integrating multiple techniques to enhance fraud detection capabilities, offering a robust framework that is less reliant on labelled datasets compared to traditional models, and addressing the challenges posed by high-dimensional multivariate data.

5. Conclusion

In conclusion, this study presents a comprehensive multivariate approach for fraud detection that effectively combines Principal Component Analysis (PCA), Isolation Forest, and KMeans clustering to address the complexities of high-dimensional transactional data. The integration of these techniques not only enhances the accuracy of detecting fraudulent activities but also allows for greater flexibility and adaptability in the face of evolving fraud tactics. Our findings indicate that this approach outperforms traditional rule-based and supervised machine learning methods by reducing false positive rates and minimizing reliance on extensive labeled datasets. The proposed methodology demonstrates significant potential for real-world applications in the financial sector, where rapid and reliable fraud detection is paramount. Future work could explore the incorporation of additional machine learning algorithms and deep learning techniques to further improve detection capabilities and robustness. Additionally, the application of this methodology to other domains experiencing fraud, such as insurance and e-commerce, could yield valuable insights and contribute to broader advancements in the field of anomaly detection. Overall, this research lays the groundwork for developing more sophisticated fraud detection systems that leverage the power of multivariate data analysis.

References

[1] Chandola, Varun, Arindam Banerjee, and Vipin Kumar. "Anomaly detection: A survey." ACM computing surveys (CSUR) 41.3 (2009): 1-58.

[2] Pang, Guansong, et al. "Deep learning for anomaly detection: A review." ACM computing surveys (CSUR) 54.2 (2021): 1-38.

[3] Chalapathy, Raghavendra, and Sanjay Chawla. "Deep learning for anomaly detection: A survey." *arXiv preprint arXiv:1901.03407* (2019).

[4] Xu, Xiaodan, Huawen Liu, and Minghai Yao. "Recent progress of anomaly detection." *Complexity* 2019.1 (2019): 2686378.

[5] Fernandes, Gilberto, et al. "A comprehensive survey on network anomaly detection." Telecommunication Systems 70 (2019): 447-489.
[6] Han, Songqiao, et al. "Adbench: Anomaly detection benchmark." Advances in Neural Information Processing Systems 35 (2022): 32142-32159.
[7] Nassif, Ali Bou, et al. "Machine learning for anomaly detection: A systematic review." Ieee Access 9 (2021): 78658-78700.
[8] Schmidl, Sebastian, Phillip Wenig, and Thorsten Papenbrock. "Anomaly detection in time series: a comprehensive evaluation." Proceedings of the VLDB Endowment 15.9 (2022): 1779-1797.
[9] Koren, Oded, Michal Koren, and Or Peretz. "A procedure for anomaly detection and analysis." Engineering Applications of Artificial Intelligence 117 (2023): 105503.
[10] Xia, Xuan, et al. "GAN-based anomaly detection: A review." Neurocomputing 493 (2022): 497-535.
[11] Bergmann, Paul, et al. "The MVTec anomaly detection dataset: a comprehensive real-world dataset for unsupervised anomaly detection." International Journal of Computer Vision 129.4 (2021): 1038-1059.
[12] Pang, Guansong, Chunhua Shen, and Anton Van Den Hengel. "Deep anomaly detection with deviation networks." Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining. 2019.
[13] Pang, Guansong, Chunhua Shen, and Anton Van Den Hengel. "Deep anomaly detection with deviation networks." Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining. 2019.
[14] Xie, Kun, et al. "On-line anomaly detection with high accuracy." IEEE/ACM transactions on networking 26.3 (2018): 1222-1235.
[15] Ahmad, Subutai, et al. "Unsupervised real-time anomaly detection for streaming data." Neurocomputing 262 (2017): 134-147.
[16] He, Shilin, et al. "Experience report: System log analysis for anomaly detection." 2016 IEEE 27th international symposium on software reliability engineering (ISSRE). IEEE, 2016.
[17] Hasan, Sakib, et al. "Frequency Translation and Filtering Techniques in Baseband Conversion." 2024 7th International Conference on Electronics Technology (ICET). IEEE, 2024.
[18] Hasan, Sakib, et al. "Neural Network-Powered License Plate Recognition System Design." Engineering 16.9 (2024): 284-300.
[19] Jannat, Syeda Fatema, et al. "AI-Powered Project Management: Myth or Reality? Analyzing the Integration and Impact of Artificial Intelligence in Contemporary Project Environments." International Journal of Applied Engineering & Technology 6.1 (2024): 1810-1820.
[20] Sunny, Md Nagib Mahfuz, et al. "Predictive Healthcare: An IoT-Based ANFIS Framework for Diabetes Diagnosis." Engineering 16.10 (2024): 325-336.
[21] Sunny, Md Nagib Mahfuz, et al. "Optimizing Healthcare Outcomes through Data-Driven Predictive Modeling." Journal of Intelligent Learning Systems and Applications 16.4 (2024): 384-402.
[22] Kavitha, M., et al. "Machine learning techniques for anomaly detection in smart healthcare." *2021 Third International Conference on Inventive Research in Computing*

Applications (ICIRCA). IEEE, 2021.
[23] Rathod, Viraj, Chandresh Parekh, and Dharati Dholariya. "AI & ML Based Anamoly Detection and Response Using Ember Dataset." 2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO). IEEE, 2021.
[24] Gupta, Sushil Kumar, et al. "Anamoly Detection in Very Large-Scale System using Big Data." 2022 International Conference on Knowledge Engineering and Communication Systems (ICKES). IEEE, 2022.
[25] Rao, B. Narendra Kumar, et al. "ML Approaches to Detect Email Spam Anamoly." 2022 International Conference on Data Science, Agents & Artificial Intelligence (ICDSAAI). Vol. 1. IEEE, 2022.
[27] Upadhyaya, Akanksha, Vinod Shokeen, and Garima Srivastava. "Analysis of counterfeit currency detection techniques for classification model." 2018 4th International Conference on Computing Communication and Automation (ICCCA). IEEE, 2018.
[28] Pareek, Manoj, et al. "Anamoly Detection in Very Large Scale System using Big Data." 2022 International Conference on Knowledge Engineering and Communication Systems (ICKES). 2023.
[29] Arvind, Siddapuram, et al. "Research Article Deep Learning Regression-Based Retinal Layer Segmentation Process for Early Diagnosis of Retinal Anamolies and Secure Data Transmission through ThingSpeak." (2022).
[30] Poluri, Sumanth Reddy, Venkata Krishna Reddy Tiyyagura, and K. Santhi Sri. "Heart Disease Prediction Based On Machine Learning." 2023 International Conference on Signal Processing, Computation, Electronics, Power and Telecommunication (IConSCEPT). IEEE, 2023.
[31] Shahul Kshirsagar, P. G., and S. R. Yadav. "Intrusion Detection Systems By Anamoly-Based Using Neural Network."
[32] Rajeev, Haritha, and Uma Devi. "Detection of credit card fraud using isolation forest algorithm." *Pervasive Computing and Social Networking: Proceedings of ICPCSN 2021*. Springer Singapore, 2022.
[33] Togbe, Maurras Ulbricht, et al. "Anomaly detection for data streams based on isolation forest using scikit-multiflow." Computational Science and Its Applications–ICCSA 2020: 20th International Conference, Cagliari, Italy, July 1–4, 2020, Proceedings, Part IV 20. Springer International Publishing, 2020.
[34] Singh, Aditi, et al. "Design and Implementation of Different Machine Learning Algorithms for Credit Card Fraud Detection." *2022 International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)*. IEEE, 2022.