Quadratic multivariate linear regressive distributed proximity feature engineering for cybercrime detection in digital fund transactions with big data

Arul Jeyanthi Paulraj[1], Balaji Thalaimalai[1,2]
[1]Department of Computer Applications and School of Management, Madurai Kamaraj University, Madurai, India
[2]Post Graduate Department of Computer Science, Government Arts College, Melur, India

Article Info

Article history:

Received May 7, 2024
Revised Aug 9, 2024
Accepted Aug 20, 2024

Keywords:

Adaptive Ziggurat synthetic sampling
Big data
Cybercrime detection
Digital fund transactions
Quadratic multivariate linear regression approach
Sokal–Michener's distributed proximity feature engineering

ABSTRACT

Digital fund transactions involve the electronic transfer of funds between parties through digital channels such as online banking platforms, mobile applications, and electronic payment systems. However, the rapid advancement of digital transactions has also directed cybercriminals to exploit vulnerabilities, engaging in money laundering and other illegal activities, resulting in substantial financial losses. The improve accuracy of cybercriminal detection by lesser time consumption, a novel technique called quadratic multivariate linear regressive distributed proximity feature engineering (QMLRDPFE) is developed. The proposed QMLRDPFE technique comprises two primary steps namely data preprocessing and feature engineering. Analyzed results prove that the QMLRDPFE technique outperforms existing methods in attaining superior accuracy and precision. Furthermore, QMLRDPFE method shows effective in reducing time utilization and space complexity for fraudulent transaction detection compared to existing approaches. Results to provide effective in reducing time utilization and space complexity for fraudulent transaction detection than the conventional methods.

Corresponding Author:

Arul Jeyanthi Paulraj
Department of Computer Applications and School of Information Technology, Madurai Kamaraj University
Madurai, Tamil Nadu, India
Email: jeyanthijayabal@gmail.com

1. INTRODUCTION

Digital payment schemes are further popular due to the increasing usage of smartphones, magnetize attention of fraudsters. A fraud detection framework based on XGBoost with random under-sampling (RUS+XGBoost) was developed by Hajek *et al.* [1] with the aim of improving fraud detection systems during mobile payment transactions. The hybridization of competitive swarm optimization as well as deep convolutional neural network (CSO-DCNN) was developed by Karthikeyan *et al.* [2] to enhance accuracy of fraudulent transaction detection. A Bayesian optimization method was developed by Hashemi *et al.* [3] for credit card fraud recognition with weight-tuning hyperparameters to mention the problem of unbalanced data while consuming lesser memory and time. Exploratory analysis and machine learning (ML) methods were designed by Moreira *et al.* [4] for predicting fraud within the banking system.

Random forest (RF) model was presented in [5] to classify online credit card transactions as fraudulent. A genetic algorithm (GA)-based feature selection technique incorporated by ML methods was designed [6] with the aim of credit card fraud detection. A logistic regression method was designed [7] to forecast transaction flagged as not during mobile cash transmits. A personalized alarm method was introduced in [8] to distinguish frauds within online fund transfers by utilizing sequence pattern mining based on users' normal transaction log files. Statistical and machine learning models were introduced in [9] for payment card fraud detection. Hybrid method combining bagging and boosting ensemble classifiers was developed in [10] for credit card fraud recognition, resulting in higher accuracy.

For detecting internet financial deception, Intelligent and dispersed big data method was developed in [11]. An unsupervised ML method was designed by Hanae [12] for detecting transactional fraud through behavioral analysis. Back propagation neural network (BPNN) model was designed Xiong *et al.* [13] for internet financial fraud identification. XGBoost and light gradient boosting machine (LGBM) methods were developed by Hsin *et al.* [14] to achieve improved fraud recognition out comes through eliminating noisy features and addressing data imbalance problem. An intelligent sampling and self-supervised learning method was developed by Chen *et al.* [15] to accurately identify credit card transactions by extracting spatial and temporal features.

For enhancing accuracy of fraud detection by balancing the majority and minority classes, dual autoencoders generative adversarial network was developed in study [16]. Hybridization of bio-inspired optimization method as well as support vector machine (SVM) was developed in [17] to enhance accuracy of credit card transaction detection. A neural network-based feature extraction method was designed in [18] that learns feature for fraud classification task. Spatio-temporal attention graph neural network (STAGN) was introduced in [19]. Credit card deception recognition method was introduced in [20]. A deep convolutional neural network (CNN) model was designed in [21] to perceive anomalies as of usual patterns created through competitive swarm optimization. Leveraging ML as well as big data analytics was performed in [22]. In study [23], big data-driven banking operations were introduced into accessibility of additional data improved difficulty of service administration as well as producing fierce competition, and so on. telecommunication network fraud depend on big data for killing pigs and plates was examined in [24]. In study [25], specifics and patterns of cybercrime were designed to examine the international community and a number of states in combating cybercrime in the field of payment processing.

The main contribution of this proposed QMLRDPFE method is follow as:

- To improve accuracy of cybercrime detection in digital fund transactions with big data, quadratic multivariate linear regressive distributed proximity feature engineering (QMLRDPFE) method is developed depend on preprocessing as well as feature engineering.
- To minimize time for fraudulent activities detection, QMLRDPFE method performs data preprocessing. The quadratic multivariate linear regression is applied for determining the missing data. The Ziggurat synthetic sampling method to solve the data imbalance.
- The QMLRDPFE technique utilizes Sokal–Michener's distributed proximity feature engineering for minimizing dimensionality of database by selecting significant features.
- Finally, experimental assessment is conducted to calculate performance of QMLRDPFE method in comparison to conventional methods.

The problem statement of our work is provided as: With advancements in machine learning, different algorithms have been enhanced to conclude whether transactions in digital systems are fraudulent or not. Convenience also brings an increased risk of cybercrime, as fraudsters exploit vulnerabilities in digital systems. The model lacks in providing improved accuracy in big data applications for fraud detection systems. The method of CSO-DCNN is failed to utilize preprocessing techniques to mention problem of uneven or unbalanced information. To overcome these issues, our proposed QMLRDPFE technique is improving the accuracy of cybercriminal detection with lesser time consumption in digital fund transactions with big data.

Manuscript is structured to five parts as pursue: Section 2 appraisal literature review. QMLRDPFE method is explained in section 3. Section 4 provides experimental setup and gives explanation of database. Comparative analyses of dissimilar parameters are given in section 5. Lastly, section 6 provides a conclusion.

2. METHOD

Fraud detection and prevention in fund transactions are crucial aspects of the modern financial system, as they help avert monetary losses as well as sustain customer trust. Therefore, financial schemes are dependable for guarantying the safety and security of their customers' funds. An efficient system is required for preventing fraud detection during digital fund transactions. In this section, a novel technique called QMLRDPFE is introduced for accurate fraud detection in digital fund transactions with minimal time consumption. Figure 1 illustrates structural design diagram of QMLRDPFE technique for accurate detection of fraudulent transactions or cybercrimes. Effective fraudulent transaction detection techniques include data acquisition, preprocessing, and feature engineering.

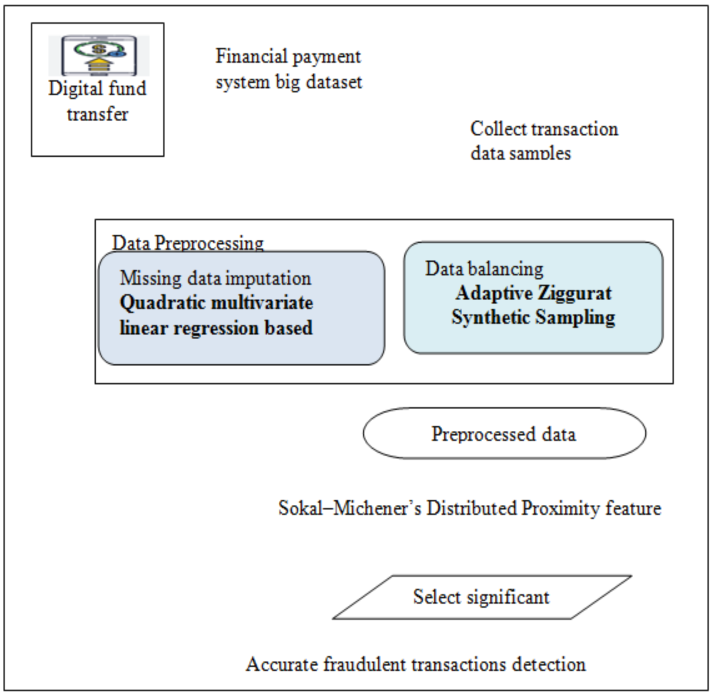
Digital fund transfer
Financial payment system big dataset
Collect transaction data samples
Data Preprocessing
Missing data imputation
Quadratic multivariate linear regression based
Data balancing
Adaptive Ziggurat Synthetic Sampling
Preprocessed data
Sokal–Michener's Distributed Proximity feature
Select significant
Accurate fraudulent transactions detection

Figure 1. Architecture of proposed QMLRDPFE technique

2.1. Data acquisition

It involves gathering relevant transaction information from different resources such as transaction logs, user profiles, or network traffic. This process utilizes the financial payment system dataset, which includes several log files totaling 594,643 records. Features namely age, customer information, transaction amount, source of transaction, target, types of transaction and labels are employed for data analysis. Depending on analysis, fraudulent activity or normal activities are identified.

2.2. Data preprocessing

It is vital part in data analysis which includes cleaning, transforming, organizing raw information to appropriate format for ensuing study and modeling. Initially, large numbers of transaction information are gathered as of datasets. However, this raw information frequently includes missing values, inconsistencies, and imbalances. To handle these issues, the proposed QMLRDPFE performs data preprocessing, which includes two main tasks such as handling missing data and addressing data imbalance problems.

2.2.1. Quadratic multivariate linear regression

Missing data refers to dearth of values in a particular column of the dataset. These missing data significantly impact the analyses of accurate fraudulent transactions defection. Therefore, handling missing data is important to ensure accurate and reliable outcomes as of big data analysis. The proposed QMLRDPFE technique utilizes the quadratic multivariate linear regression for handling missing data in a given dataset.

Quadratic multivariate linear regression is the ML method employed to predict missing values based on multiple available data. Multivariate data indicates multiple available data in the dataset used for finding the missing values. Let us assume input dataset 'Ds'as well as formulated in matrix,

$$IM = \begin{bmatrix} a_1 & a_2 & \dots & a_m \\ DP_{11} & DP_{12} & \dots & DP_{1n} \\ DP_{21} & DP_{22} & \dots & DP_{2n} \\ \vdots & \vdots & \dots & \vdots \\ DP_{m1} & DP_{m2} & \dots & DP_{mn} \end{bmatrix} \quad (1)$$

where, IM indicates an input data matrix, each column indicates a number of features $a_1, a_2, a_3, \dots a_m$, each row indicates a number of data samples or instances $DP_1, DP_2, DP_3, \dots DP_n$ respectively. Quadratic linear regression is used to measure the relationship between the independent variables *i.e.* data samples DP_n is modeled as (2),

$$Q = \delta_0 + \delta_1 DP_1 + \delta_2 {DP_2}^2 + \cdots \delta_n {DP_n}^n + \epsilon \quad (2)$$

where, Q denotes an output of quadratic linear regression, $DP_1, DP_2, DP_3, \dots DP_n$ denotes a number of data samples or instances, $\delta_0, \delta_1, \delta_3, \dots \delta_n$ denotes a coefficients of the quadratic regression equation, ϵ indicates the error term which minimizes the sum of squared variation among examined Q_a as well as forecasted values Q_p.

$$\epsilon = arg\, min \left(Q_a - Q_p\right)^2 \quad (3)$$

The quadratic function involves finding the values of the coefficients that minimize *i.e.* $argmin$ the least absolute deviation between the observed Q_a and predicted values Q_p. In this way, these proposed an imputation technique effectively handles all missing values in the given dataset.

2.2.2. Adaptive Ziggurat synthetic sampling for handle imbalance data

Data imbalance is addressed where allocation of classes in database not even. In dataset, one or more classes have significantly fewer instances than others. This imbalance poses challenges for ML methods as become biased toward mainstream class, out come at deprived results on minority class. To solve this issue, adaptive Ziggurat synthetic sampling technique is employed in the proposed QMLRDPFE to generate synthetic data for minority class, aiming to balance class allocation in dataset. This process is particularly useful for improving the accuracy of fault detection in digital fund transactions.

Imbalanced data is handled by applying adaptive Ziggurat synthetic sampling for generating number of information samples at minority class. Initially, define target amount of synthetic data samples needs to be generated for the minority class as (4).

$$S = \ MxC_{DP} - \ MnC_{DP} \quad (4)$$

where, S denotes target amount of synthetic information samples needs to create, MxC_{DP} indicates a majority counts of data samples, represents a minority counts of data samples in the dataset.

After finding the counts to generate synthetic data samples, the sampling process is executed. adaptive Ziggurat synthetic sampling is a method used for generating data samples from a Gaussian probability distribution of the other data samples in the dataset. It is an efficient method compared to other methods. First, consider the random numbers 'R' from 0 to 1 *i.e.* [0, 1] since Ziggurat synthetic sampling utilizes the Gaussian probability distribution.

$$R = \{R_1, R_2, R_3, \dots\ R_i\} \quad (5)$$

Secondly, initialize the k number of layers in the Gaussian probability distribution as shown in Figure 2. Figure 2 illustrates the layer segmentation in Gaussian distribution where k indicates a number of layers and red point indicates a boundary of the layer b_1, b_2, b_3 respectively. For each random number, then compute the following function Q,

$$Q = R_J * \ B_K \quad (6)$$

where, R_j denotes a random number, b_k indicates a boundary of the layer. After that, the probability density function is computed with '0' mean and deviation '1'.

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left[\frac{(Q-\mu)^2}{2\sigma^2}\right] \quad (7)$$

By applying '0' mean (μ) and deviation (σ) '1', the above equation becomes written as (8),

$$f(x) = \ exp\left[\frac{(Q)^2}{2}\right] \quad (8)$$

The process then verifies that the computed 'Q' falls within the specified range the $f(x)$. Then the value of Q is selected as a synthetic data sample. Otherwise, it rejects the generated samples and repeats the above process until the target amount of synthetic information samples is reached. Like this, data imbalance problems are handled in the proposed techniques.

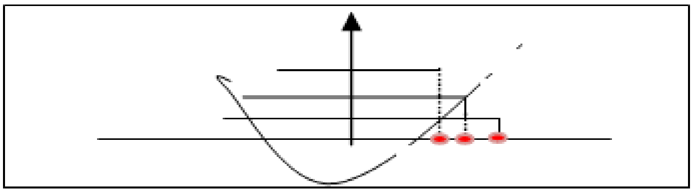

Figure 2. Layers of Gaussian distribution

Algorithm 1 given clearly describes data preprocessing to reduce time utilization of financial fraud prediction during the digital fund transaction. Initially, a number of data samples are gathered as of dataset. Next, missing values are recognized by applying quadratic linear regression. Once missing values are filled, the issue of data imbalance is addressed. Firstly, the target number of synthetic data samples is determined. Then, random numbers are generated. Subsequently, it multiplied through a predefined boundary. Following this; the probability density function is estimated with zero mean and one standard deviation. Each estimated data point is validated against the probability density function. If its value is lesser than that of the probability density function, it is selected as a synthetic data sample. Otherwise, the data sample is rejected. This process continues until the target number of synthetic data samples is reached.

Algorithm 1. Data pre-processing

```
Input: Dataset '$Ds$', features $a_1, a_2, a_3, \dots a_m$, data samples or instances $DP_1, DP_2, DP_3, \dots DP_n$
Output: Pre-processed dataset
 Begin
1. For each dataset '$Ds$' with features '$a$'
2.     Formulate input vector matrix '$IM$' using (1)
3.     If missing value in dataset then
4.        Apply quadratic linear regression using (2)
5.         Fill the value to the respective missing column
6.    End if
7.    Find number of target data samples needs to be generated using (3)
8.     Define the random numbers using (5)
9.   Define the numbers layers '$k$' and boundary '$b$' using
10.      Measure the product of the random numbers and boundary using (6)
11.         compute the probability density function with zero mean and deviation using (8)
12.  if ($Q < f(x)$) then
13.          Selected as a synthetic data samples
14.      else
15:         Reject the data samples
16.    end if
17.        Go to step 8
18.     Obtain the number of synthetic data samples
19. Return (balanced dataset)
20.  End for
 End
```

2.3. Sokal–Michener's distributed proximity feature engineering

With the balanced data set, the feature engineering process is executed for dimensionality reduction. Dimensionality reduction is a technique to minimize the number of features within a big dataset. Big datasets include a more number of features which causes increased computational complexity and challenges in achieving accurate classification. To mention this issue, the Sokal–Michener's distributed proximity feature engineering method is developed in QMLRDPFE for dimensionality reduction by selecting the most significant features. Through the identification of significant features, this approach enhances the accuracy of cybercrime detection, specifically in classifying the fraudulent activities within digital fund transactions.

Figure 3 flow process of the Sokal–Michener's distributed proximity feature engineering for accurate fraudulent activities detection. Let us consider the number of features $a_1, a_2, a_3, \dots a_m$ distributed in the given dataset. Proximity refers to the degree of closeness or similarity between two features in a database. Afterward using Sokal–Michener's for determining similarity between features.

$$FP = SM\,(a_i, a_j) \tag{9}$$

where, FP denotes a feature proximity, $SM\,(a_i, a_j)$ indicates a Sokal–Michener's similarity. It is measured as (10),

$$SM\ (a_i, a_j) = 1 - \frac{|a_i \,\Delta\, a_j|}{m} \tag{10}$$

where, SM indicates a Sokal–Michener's similarity, $a_i \Delta a_j$ denotes a difference between the two features, m denotes a total number of features. The Sokal–Michener's similarity provides the outcomes ranges from 0 to 1.

$$Y = \begin{cases} SF,\ SM\left(a_i, a_j\right) > T \\ IF,\ SM\left(a_i, a_j\right) < T \end{cases} \tag{11}$$

where Y denotes an output function, T indicates a threshold for similarity coefficient $SM\left(a_i, a_j\right)$ results. If the coefficient $SM\left(a_i, a_j\right)$ exceeds the threshold, the feature is termed as significant feature (SF). Otherwise, it is termed as insignificant feature (IF). Finally, the significant features are selected for accurate fraudulent transaction detection and other features are removed from the dataset. The algorithm for Sokal–Michener's distributed proximity feature engineering is given.

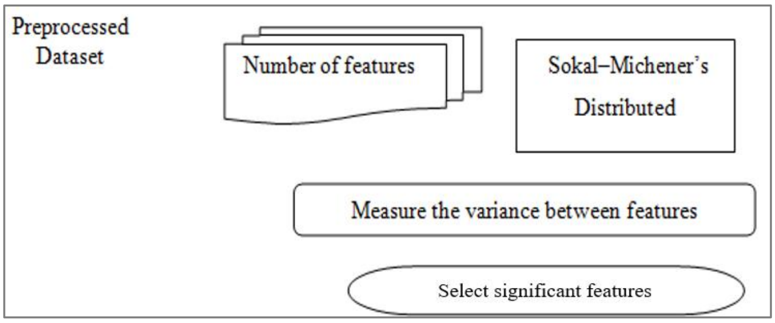
Preprocessed Dataset
Number of features
Sokal–Michener's Distributed
Measure the variance between features
Select significant features

Figure 3. Flow process of Sokal-Michener's distributed proximity feature engineering

Algorithm 2 describes the process of significant feature selection using Sokal–Michener's distributed proximity feature engineering technique for improving fraudulent transaction detection while reducing time utilization. The preprocessed dataset comprises several features used as the input. Subsequently, feature proximity is computed between the features based on Sokal–Michener's similarity measure. This similarity measure distinguishes the significant and insignificant features by setting the threshold within the dataset. Finally, important features are chosen to improve accuracy of fraudulent detection in digital fund transactions.

Algorithm 2. Sokal–Michener's distributed proximity feature engineering

```
Input: Preprocessed datasets '$Ds$', features $a_1, a_2, a_3, \dots a_m$,
       data samples or instances $DP_1, DP_2, DP_3, \dots DP_n$
Output: Select relevant features
 Begin
 Collect the preprocessed dataset as input
        For each feature '$a_i$'
               Measure the proximity using (9)
               Measure the Sokal-Michener's similarity '$SM\ (a_i, a_j)$'
          if ($SM\ (a_i, a_j) > T$) then
               Features are identified as significant
          else
               Features are identified as insignificant
          End if
               Select the significant features and remove other features
        end for
 End
```

3. EXPERIMENTAL SCENARIO

Experimental assessment of QMLRDPFE technique and existing XGBoost-based fraud detection framework [1] and CSO+DCNN [2] are executed by Python coding. To carry out experiment, financial payment system dataset is collected as of Kaggle dataset [26]. Major objective of this database is employed to identify fraudulent transactions and normal payments. This dataset includes a several log files that include 594,643 records. Dataset includes 10 features such as step, customer, age, gender, zipcodeOri, and so on. In order to conduct the experiment, the number of data samples is considered in the ranges from 10,000 to 100,000.

3.1. Implementation details

In this study, we developed a novel technique called quadratic multivariate linear regressive distributed proximity feature engineering (QMLRDPFE) is developed to enhance the accuracy of cybercriminal detection with minimum time consumption.

- The QMLRDPFE method comprises two primary steps namely data preprocessing and feature engineering.
- We compared our QMLRDPFE technique compared to existing XGBoost-based fraud detection framework [1] and CSO+DCNN [2] using financial payment system dataset to validate the results.
- The database contains payments from different customers made at dissimilar time periods as well as through diverse amounts. Main aim of this database is used to detect the fraudulent transactions and normal payments
- Initially the preprocessing is carried out, involving two key processes namely handling missing data and balancing the dataset. The missing information depends on multiple available data as well as imputed information is to reduce least absolute deviation.
- After that the feature engineering technique, selecting the most relevant features. During the identification of significant features, specifically in classifying the fraudulent activities within digital fund transactions in this dataset.

4. PERFORMANCE COMPARISION ANALYSIS

In this section, performance of the proposed QMLRDPFE technique and existing RUS+XGBoost [1] and CSO+DCNN [2] are assessed with various metrics, including accuracy, precision, execution time and space complexity.

- Detection accuracy: It is defined to ratio of accurately detecting fraudulent transactions and normal transactions as of total number of data. It is computed as (12),

$$DAcc = \left(\frac{Tps+Tng}{Tps+Tng+Fps+Fng}\right) * 100 \tag{12}$$

  where, $DAcc$ indicates a detection accuracy, Tps indicates true positive, Tng symbolize true negative, Fps denotes false positive, and Fng indicates false negative. It is calculated in percentage (%).
- Precision: It is defined as ratio of detecting fraudulent transactions and normal transactions. It is computed as (13),

$$Prs = \left(\frac{Tps}{Tps+Fps}\right) \tag{13}$$

  where, Prs denotes a precision, Tps denotes the true positive, and Fps represents the false positive.
- Detection time: It is measured as the amount of time consumed by algorithm for detecting the fraudulent transactions and normal transactions. The time is computed as (14),

$$DT = \sum_{i=1}^{n} DP_i * TM(D) \tag{14}$$

  where, DT denotes a detection time depend on data samples DP_i as well as actual time utilized in detecting the fraudulent transactions and normal transactions denoted by $TM(D)$. It is calculated in milliseconds (ms).
- Space complexity: It is calculated as amount of memory space utilized through method for detecting the fraudulent transactions and normal transactions. The Space complexity is computed as (15),

$$SC = \sum_{i=1}^{n} DP_i * Mem(D) \tag{15}$$

  where, SC denotes a space complexity depend on data samples DP_i and memory space utilized at detecting the fraudulent transactions and normal transactions denoted by $Mem(D)$. It is calculated in kilobytes (kB).

Table 1 given above illustrates performance comparison of detection accuracy of fraudulent transactions and normal payments using three methods namely QMLRDPFE technique and existing RUS+XGBoost [1] and CSO+DCNN [2]. Among the three techniques, performance of QMLRDPFE method is improved than the conventional techniques. For example, measuring 10,000 data samples computing detection accuracy, QMLRDPFE method attained accuracy of 90%. As well, $DAcc$ of conventional [1], [2] was 86% and 982%, respectively. For each method, ten different outcomes are examined. The observed outcomes are compared. Overall comparative study denotes which $DAcc$ of QMLRDPFE method enhanced by 5% and 3% than the study [1], [2]. This is due to utilizing Sokal–Michener's distributed proximity feature engineering method is developed for dimensionality reduction by selecting the significant features. Depend on accurately performs cybercrime detection, by distinguishing the fraudulent activities or normal during the digital fund transactions.

Figure 4 depicts a comparison of precision. Three methods, namely QMLRDPFE technique, existing RUS+XGBoost [1], and CSO+DCNN [2], are utilized for calculating precision. Outcomes demonstrate which QMLRDPFE technique achieves superior Prs than conventional techniques. Observed results of QMLRDPFE method are compared to existing methods. Overall comparison reveals that the precision performance in accurately detecting fraudulent activities during digital fund transactions is enhanced by 6% and 3% than the [1], [2] when applying the QMLRDPFE technique. To achieve this improved performance, the QMLRDPFE technique utilizes Sokal–Michener's distributed proximity feature engineering technique for selecting target features, thereby enhancing detection through improved Tps and minimizing Fps outcomes during fraudulent transaction detection.

Given above depicts the performance comparison of detection time by QMLRDPFE technique, existing RUS+XGBoost [1], and CSO+DCNN [2]. Performance of for every three techniques obtain enhanced as enhancing number of data samples. Especially, for QMLRDPFE technique is minimized than the [1], [2]. Let us assume initial iteration with 10,000 data samples, where DT for QMLRDPFE method was likewise, time utilization for [1], [2] respectively. The obtained overall results of QMLRDPFE method are compared to outcomes of conventional techniques. The comparison outcomes denotes which performance of detection time using QMLRDPFE technique is significantly reduced by 13% and 7% than the study [1], [2]. This is owing to QMLRDPFE technique performed the data preprocessing and feature selection process. In data preprocessing, missing information is determined by applying quadratic multivariate linear regression approach. Data imbalance problem also solved through the adaptive Ziggurat synthetic sampling technique to create synthetic data samples. The target feature selection also minimizes time consumption of fraudulent transaction detection.

Table 1. Comparison of detection accuracy

Number of data samples	Detection accuracy (%)		
	QMLRDPFE	RUS+XGBoost	CSO+DCNN
10000	90	87	88.5
20000	91.22	88.52	89.85
30000	90.33	86.23	87.52
40000	90.88	85.99	87.78
50000	92.12	86.89	89.52
60000	91.05	86.74	88.74
70000	92.05	85.52	87.22
80000	91.5	86.56	88.56
90000	92.2	87.52	89.5
100000	91.88	85.98	87.22

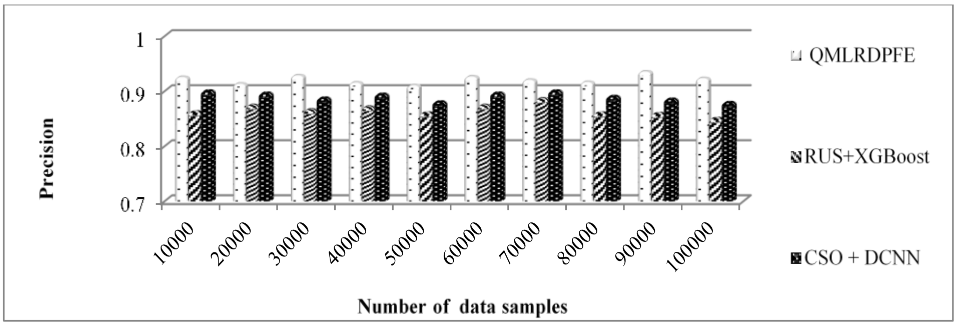

Figure 4. Performance comparison of Prs

Table 2 and Figure 5 depicts the performance comparison of detection time by QMLRDPFE technique, existing RUS+XGBoost [1], and CSO+DCNN [2]. Performance of for every three techniques obtain enhanced as enhancing number of data samples. Especially, for QMLRDPFE technique is minimized than the study [1], [2]. Let us assume initial iteration with 10,000 data samples, where DT for QMLRDPFE method was likewise, time utilization for [1], [2] respectively. The obtained overall results of QMLRDPFE method are compared to outcomes of conventional techniques. The comparison outcomes denotes which performance of detection time using QMLRDPFE technique is significantly reduced by 13% and 7% than the study [1], [2]. This is owing to QMLRDPFE technique performed the data preprocessing and feature selection process. In data preprocessing, missing information is determined by applying quadratic multivariate linear regression approach. Data imbalance problem also solved through the adaptive Ziggurat synthetic sampling technique to create synthetic data samples. The target feature selection also minimizes time consumption of fraudulent transaction detection. Table 3 denotes comparison of space complexity among the following algorithms like QMLRDPFE, RUS+XGBoost and CSO+DCNN. Sample data set values range from 10,000 to 100,000. It shows the space complexity of the dataset values.

Table 2. Comparison of detection time

Number of data samples	Detection time (ms)		
	QMLRDPFE	RUS+XGBoost	CSO+DCNN
10000	43	53	48
20000	50	60	54
30000	63	72	69
40000	72	88	80
50000	80	91	85
60000	88.8	96	91.2
70000	93.1	101.5	96.6
80000	96.8	112	108
90000	106.2	118.8	112.5
100000	116	130	125

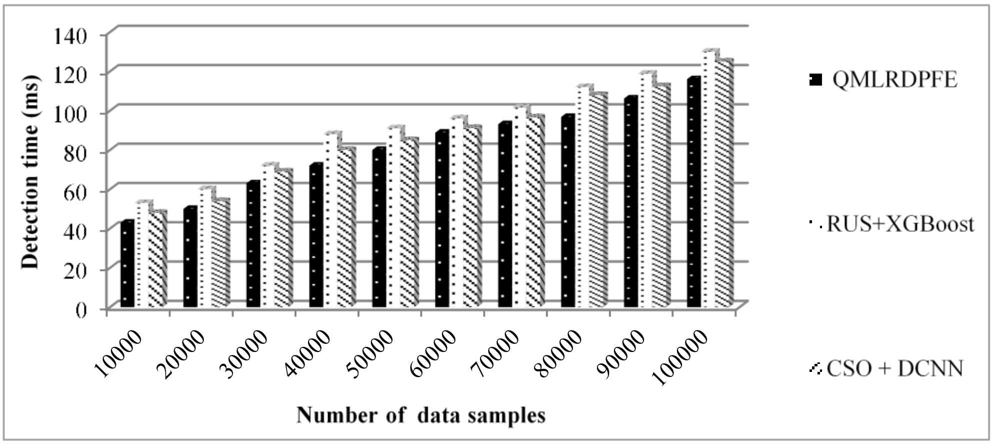

Figure 5. Performance comparison of detection time

Figure 6 depicts result outcomes of space complexity versus number of data samples extracted. While the number of information samples increases, the space complexity of every three methods gradually increases. Notably, the space complexity for the QMLRDPFE method is significantly minimized than the [1], [2]. Let's consider the results from the first iteration with 10,000 data samples. The space complexity for the QMLRDPFE technique was used to calculate space complexity [1], [2] respectively. Subsequently, the overall outcomes of QMLRDPFE technique are compared to conventional techniques. The average results demonstrate that performance of space complexity is minimized by 20% and 9% than the existing RUS+XGBoost [1] and CSO+DCNN [2], respectively. This reduction in space complexity is achieved due to the QMLRDPFE techniques performs the dimensionality reduction through Sokal–Michener's distributed proximity feature engineering technique. This approach selects significant features while removing others from the dataset, thereby minimizing storage space in big data analysis.

Table 3. Comparison of space complexity

Number of data samples	Space complexity (kB)		
	QMLRDPFE	RUS+XGBoost	CSO+DCNN
10000	320	420	380
20000	378	462	433
30000	433	510	485
40000	457	546	505
50000	501	612	532
60000	522	675	568
70000	548	724	610
80000	593	763	633
90000	635	812	687
100000	687	824	736

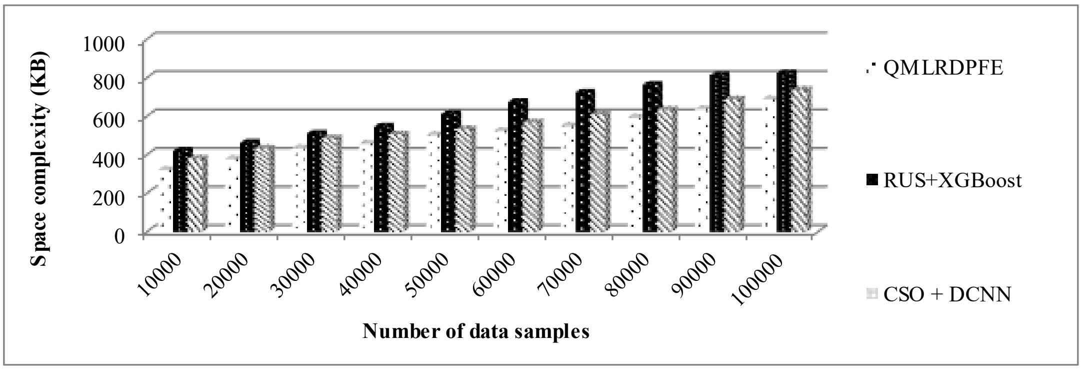

Figure 6. Performance comparison of space complexity

5. CONCLUSION

In this manuscript, a new technique called QMLRDPFE is designed for cybercrime detection in digital fund transactions. QMLRDPFE technique includes data preprocessing in the first stage to arrange the dataset properly by filling in missing data before utilizing ML method. Following this, dimensionality reduction is implemented by Sokal–Michener's distributed proximity feature engineering technique for fraudulent transaction detection with higher accuracy and Prs. A comprehensive experimental assessment is performed with detection accuracy, precision, detection time, and space complexity. The analyzed results prove which QMLRDPFE method is better than conventional methods in achieving higher accuracy and precision. In addition, QMLRDPFE method proves more effective at reducing time utilization and space complexity for fraudulent transaction detection than the conventional methods.

REFERENCES

[1] P. Hajek, M. Z. Abedin, and U. Sivarajah, "Fraud detection in mobile payment systems using an XGBoost-based framework," *Information Systems Frontiers*, vol. 25, no. 5, pp. 1985–2003, Oct. 2023, doi: 10.1007/s10796-022-10346-6.

[2] T. Karthikeyan, M. Govindarajan, and V. Vijayakumar, "An effective fraud detection using competitive swarm optimization based deep neural network," *Measurement: Sensors*, vol. 27, Jun. 2023, doi: 10.1016/j.measen.2023.100793.

[3] S. K. Hashemi, S. L. Mirtaheri, and S. Greco, "Fraud detection in banking data by machine learning techniques," *IEEE Access*, vol. 11, pp. 3034–3043, 2023, doi: 10.1109/ACCESS.2022.3232287.

[4] M. Â. L. Moreira *et al.*, "Exploratory analysis and implementation of machine learning techniques for predictive assessment of fraud in banking systems," *Procedia Computer Science*, vol. 214, pp. 117–124, 2022, doi: 10.1016/j.procs.2022.11.156.

[5] J. K. Afriyie *et al.*, "A supervised machine learning algorithm for detecting and predicting fraud in credit card transactions," *Decision Analytics Journal*, vol. 6, Mar. 2023, doi: 10.1016/j.dajour.2023.100163.

[6] E. Ileberi, Y. Sun, and Z. Wang, "A machine learning based credit card fraud detection using the GA algorithm for feature selection," *Journal of Big Data*, vol. 9, no. 1, Dec. 2022, doi: 10.1186/s40537-022-00573-8.

[7] M. E. Lokanan, "Predicting mobile money transaction fraud using machine learning algorithms," *Applied AI Letters*, vol. 4, no. 2, Apr. 2023, doi: 10.1002/ail2.85.

[8] J. Kim, H. Jung, and W. Kim, "Sequential pattern mining approach for personalized fraudulent transaction detection in online banking," *Sustainability*, vol. 14, no. 15, Aug. 2022, doi: 10.3390/su14159791.

[9] M. Seera, C. P. Lim, A. Kumar, L. Dhamotharan, and K. H. Tan, "An intelligent payment card fraud detection system," *Annals of Operations Research*, vol. 334, no. 1–3, pp. 445–467, Mar. 2024, doi: 10.1007/s10479-021-04149-2.

[10] V. S. S. Karthik, A. Mishra, and U. S. Reddy, "Credit card fraud detection by modelling behaviour pattern using hybrid ensemble model," *Arabian Journal for Science and Engineering*, vol. 47, no. 2, pp. 1987–1997, Feb. 2022, doi: 10.1007/s13369-021-06147-9.

[11] H. Zhou, G. Sun, S. Fu, L. Wang, J. Hu, and Y. Gao, "Internet financial fraud detection based on a distributed big data approach with Node2vec," *IEEE Access*, vol. 9, pp. 43378–43386, 2021, doi: 10.1109/ACCESS.2021.3062467.

[12] A. Hanae, B. Abdellah, E. Saida, and G. Youssef, "End-to-end real-time architecture for fraud detection in online digital transactions," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 6, 2023, doi: 10.14569/IJACSA.2023.0140680.

[13] T. Xiong, Z. Ma, Z. Li, and J. Dai, "The analysis of influence mechanism for internet financial fraud identification and user behavior based on machine learning approaches," *International Journal of System Assurance Engineering and Management*, vol. 13, no. S3, pp. 996–1007, Dec. 2022, doi: 10.1007/s13198-021-01181-0.

[14] Y.-Y. Hsin, T.-S. Dai, Y.-W. Ti, M.-C. Huang, T.-H. Chiang, and L.-C. Liu, "Feature engineering and resampling strategies for fund transfer fraud with limited transaction data and a time-inhomogeneous modi operandi," *IEEE Access*, vol. 10, pp. 86101–86116, 2022, doi: 10.1109/ACCESS.2022.3199425.

[15] C.-T. Chen, C. Lee, S.-H. Huang, and W.-C. Peng, "Credit card fraud detection via intelligent sampling and self-supervised learning," *ACM Transactions on Intelligent Systems and Technology*, vol. 15, no. 2, pp. 1–29, Apr. 2024, doi: 10.1145/3641283.

[16] E. Wu, H. Cui, and R. E. Welsch, "Dual autoencoders generative adversarial network for imbalanced classification problem," *IEEE Access*, vol. 8, pp. 91265–91275, 2020, doi: 10.1109/ACCESS.2020.2994327.

[17] A. Singh, A. Jain, and S. E. Biable, "Financial fraud detection approach based on firefly optimization algorithm and support vector machine," *Applied Computational Intelligence and Soft Computing*, vol. 2022, pp. 1–10, Jun. 2022, doi: 10.1155/2022/1468015.

[18] K. Ghosh Dastidar, J. Jurgovsky, W. Siblini, and M. Granitzer, "NAG: neural feature aggregation framework for credit card fraud detection," *Knowledge and Information Systems*, vol. 64, no. 3, pp. 831–858, Mar. 2022, doi: 10.1007/s10115-022-01653-0.

[19] D. Cheng, X. Wang, Y. Zhang, and L. Zhang, "Graph neural network for fraud detection via spatial-temporal attention," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 8, pp. 3800–3813, Aug. 2022, doi: 10.1109/TKDE.2020.3025588.

[20] F. A. Ghaleb, F. Saeed, M. Al-Sarem, S. N. Qasem, and T. Al-Hadhrami, "Ensemble synthesized minority oversampling-based generative adversarial networks and random forest algorithm for credit card fraud detection," *IEEE Access*, vol. 11, pp. 89694–89710, 2023, doi: 10.1109/ACCESS.2023.3306621.

[21] S. R. Byrapu Reddy, P. Kanagala, P. Ravichandran, D. R. Pulimamidi, P. V Sivarambabu, and N. S. A. Polireddi, "Effective fraud detection in e-commerce: Leveraging machine learning and big data analytics," *Measurement: Sensors*, vol. 33, Jun. 2024, doi: 10.1016/j.measen.2024.101138.

[22] M. Hasan, A. Hoque, and T. Le, "Big data-driven banking operations: opportunities, challenges, and data security perspectives," *FinTech*, vol. 2, no. 3, pp. 484–509, Jul. 2023, doi: 10.3390/fintech2030028.

[23] G. Li and Y. Wen, "Research on the detection countermeasures of telecommunication network fraud based on big data for killing pigs and plates," *Journal of Robotics*, vol. 2022, pp. 1–11, Mar. 2022, doi: 10.1155/2022/4761230.

[24] B. Sturc, T. Gurova, and S. Chernov, "The specifics and patterns of cybercrime in the field of payment processing," *International Journal of Criminology and Sociology*, vol. 9, pp. 2021–2030, Apr. 2022, doi: 10.6000/1929-4409.2020.09.237.

[25] M. Nassereddine, G. Nassreddine, and T. ElHassan, "Electric vehicle and photovoltaic advanced roles in enhancing the financial performance of a manufacturing and commercial setup," *International Journal of Electrical and Computer Engineering*, vol. 14, no. 3, pp. 2491–2499, Jun. 2024, doi: 10.11591/ijece.v14i3.pp2491-2499.

[26] E. Lopez-Rojas, "Synthetic data from a financial payment system," *Kaggle*, 2018. https://www.kaggle.com/datasets/ealaxi/banksim1 (accessed May 07, 2024).

BIOGRAPHIES OF AUTHORS

Arul Jeyanthi Paulraj research scholar in Madurai Kamaraj University, Madurai. Assistant professor in Department of Computer Science, Ayya Nadar Janaki Ammal College, Sivakasi-626124, Virudhunagar District, Tamil Nadu, India. She has completed her M.C.A in SFR College for Women, Sivakasi and M.Phil. in Madurai Kamaraj University. She has twelve years of teaching experience. She can be contacted at email: jeyanthijayabal@gmail.com.

Balaji Thalaimalai associate professor in P.G. Department of Computer Science, Government Arts College, Melur-625106, and research supervisor in Madurai Kamaraj University, Madurai, Tamil Nadu, India. He has completed his M.C.A., M.Phil. in Alagappa University, Karaikudi, M.Tech. in Manonmaniam Sundaranar University, Tirunelveli, and Ph.D. in Madurai Kamaraj University, Madurai. He has published twenty-one papers and four books. He has twenty-four years of teaching experience. He can be contacted at email: bkmdgacm1976@gmail.com.