**International Journal of Applied Mathematics**

**Volume 38 No. 3s, 2025**

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

# OPTIMIZING CLUSTERING OF K-MEANS ALGORITHM USING PARTICLE SWARM OPTIMIZATION FOR CREDIT CARD FRAUD DETECTION

**Mona Mohammed Badai[1] and Athraa Jasim Mohammed[1]**

[1]Department of Computer Science, University of Technology - Iraq,
Baghdad, Iraq
cs.24.38@grad.uotechnology.edu.iq;
athraa.j.mohammed@uotechnology.edu.iq

**Abstract –**

Since then, online shopping has grown by leaps and bounds, and most of those digital checkouts now work with credit cards. But plastic used to solely be found in wallets; today it accompanies data packets across multiple servers. The negative is exposure: $24.26 billion in credit card fraud was written off in 2018, leaving actual consumers with a large percentage of the losses. Surveillance companies call a report fraud, and then they have to convince an angry cardholder to wait until the charges are dropped. The consumers who are too elderly to take the jump and start utilizing payment applications are always the first to feel the pain when they scroll by fewer of those pop-up boxes that keep us from seeing the terms and conditions. Supervised learning is the basis of traditional defenses. This means that all hijacked charges must be identified before a detector can be trained. Before any algorithm runs, there is a lot of work to be done in labeling, annotating, and grinding. Researchers still do all of that. The newest chorus attempts to accomplish things on its own, grouping data without pre-markers. This means it has to give up speed for certainty. Here, we look at some of the same no-label techniques to see if they may still be useful as fraud evolves. To achieve that goal, we build a hybrid architecture called PSOKClus, which combines particle-swarm optimization with K-means. The approach lets swarms change the centers of clusters on the fly, which cuts down on distance, tightness, and outlier noise in a single layer. The combination of Particle Swarm Optimization with K-means aims to address the challenge of centroid initialization, a factor often cited as a cause for diminished efficacy in fraud detection. After we have the temporary centroids that have been stabilized and hard assignments are established, the cluster-specific Interquartile Range approach looks for spots that are very different from the regular behavior of each group. A number of criteria, such as the Silhouette Score, Davies–Bouldin Index, Dunn Index, overall accuracy, and Purity, show that the PSOKClus variation beats the traditional K-means by a wide margin. The setup reduces the variance within each cluster and increases the distance between the cluster centroids. It also labels flagged transactions without using labeled samples to help shape the clusters themselves, which creates a unique profile of what confirmed fraud cases look like. This hybrid architecture reliably merges coherence, separation, and functional anomaly detection over several trials. DO NOT BREAK UP PARAGRAPHS.

KEYWORDS: Interquartile Range (IQR), Cluster, Outliers Detection, integrates Particle Swarm Optimization (PSO), K-means

Received: July 21, 2025

## Introduction

When someone commits credit card fraud, consumers and issuers lose billions of dollars every year. One clear reason why detection systems fail is the virtually total lack of balance in the court class: practically all of the good activity we can identify gets into the authorized files, while abuse warnings barely fill a little dot. But since such catastrophic transactions are so rare, most models are trained on a lot of regular transactions and don't see the few bad ones that actually matter. Rules and heuristics also fail, since thieves exploit their methods while the codec stays unchanged, leading to a surge of false flags [1], also known as false positives.

There have been so many tries to address this, each one attempting to make the excellent and poor samples the same and going in various technological ways. Some teams build synthetic records on the fly, while others connect numerous classifier outputs into an ensemble or route a claim via a single deep net [2]. People who work on optimization set settings so that the uneven border may be pushed back, while people who spot anomalies ride next to the pipelines and knock down barriers around things that seem suspicious. 2014; up2014; wherever2014; it2014; was2014; going. Investigators look closely at the titles of the columns and only choose predictors that would boost lift. They then add a layer of explainability to the engine so that investigators can understand why a transaction has been flagged. This benchmark suite is now on test beds and includes the most recent heists, thus figures reported in journals are no longer out of date as soon as the page is printed [3].

There are a lot of challenges with credit card fraud detection that make it not operate well when people are fighting over money. One of the biggest problems is that there are too many transaction records and not enough fraud samples. Almost any algorithm trained on the whole data set would filter out fraud examples, making them noise. Rule-based engines and static dashboards become useless nearly instantly since criminals change their tactics rapidly and are continually attempting new things. While smart concepts are often used in current countermeasures, very few provide a believable synthetic fraud use case, cross-merchant vertical efficacy, or flexibility against the next kind of fraud [4]. If you only look at the holes that are left, it's clear to see why this happened: Researchers still need strong algorithms that can find bad behavior in skewed data. These systems need to be able to keep up with the changing methods that people commit fraud.

Detecting fake transactions is a hard and crucial topic in finance. In RS, it talks about how to use PSO with the K-means clustering approach and the IQR method to create a good Credit Card Fraud Detection model. In which standard clustering techniques (e.g., k-means) are often susceptible to random initialization and outliers. By changing the PSO algorithm, the better placement of cluster centers increases the accuracy of detection. Using the IQR approach, you can tell the difference between legitimate transactions and clusters that include fraudulent transactions. This integration might show the flaws in conventional clustering, and the corruption detection systems could become more accurate and stable.

## Related Work

The increasing use of contactless payments and online transfers has sparked a resurgence of interest in detecting credit card fraud. The two problems of class inequality and fast-moving

Received: July 21, 2025

con artists are still puzzling to academics. Recent research endeavors have significantly depended on clustering algorithms and optimization methodologies to formulate viable answers.

CreditAccess Grameen (2021) [14] identified a cluster in debit card flows, providing fresh insights with the use of clustering methods. [7] The initial stage in their process is an Approximate Page Rank sweep that sets up the graph. From there, a merchant network is built up and put into Node2Vec. Then, conventional DBSCAN is used to make typical Euclidean clusters. This counts the output groups by risk score, which lets analysts rapidly find the right ones. The algorithm detected 145 merchants who were unusually dangerous during a live trial and found 178 new cash-out operators from the same pool. This brought the total number of entries on the watch-list to 30,586. Chinese banking regulators have already added the framework to real-time processing for this framework. This shows that it is already ready for the street.

Manda et al. Credit-card fraud detection for a highly unbalanced situation, with less than two transactions out of 100 being fraudulent, (2022) [8] The researchers evaluated five ensemble classifiers—Random Forest, AdaBoost, CatBoost, XGBoost, and LightGBM—against the task, using metrics and models. But XGBoost prevailed with an area-under-the-curve score of 0.974, while AdaBoost came in second with 0.83. During this experiment, data cleansing, trend graphing, and numerical sanity tests stayed in the middle stage, with none of them losing concentration. SMOTE-GAN, the hybrid that uses generative adversarial networks to generate even more realistic fake samples, was not on the list. The order in which cards are swiped wasn't looked at, which is a weakness that puts the toolkit open to quickly shifting frauds like skimming.

Manayi and Jebari (2022) [10] used a hybrid approach of support-vector data-description and particle-swarm optimization to identify fraudulent invoicing. The swarm technique was used to adjust the sparse-matrix parameters C and σ such that the method was able to make claims with up to 97 percent accuracy in three well-known unbalanced corpuses. Standard SVM, on the other hand, did not come close to that target.

N. Prabhakaran and R. Nedonchelian (2023) [10] have recently suggested an Online Credit-Card-Fraud Detection (CCFD) framework that is based on a continually-controlled feature-selection (OCSODL-CCFD). This method uses the ordered-closed-set-optimizer (OCSO) and deep-learning (DL) models to get rid of all but the most useful input variables for practitioners. The next phase in the CKHA of parameter fine-tuning makes the algorithm more able to deal with problems that happen in the actual world. Simulation studies demonstrate that, relative to a set of baselines, OCSODL-CCFD excels; nonetheless, the first publication fails to account for potential performance bias arising from abrupt feature drifts or overt artificiality in data, indicating a significant opportunity for additional investigation.

H. Ahmed and others A methodology to address data imbalance in credit card fraud detection by using oversampling and fuzzy C-means clustering approaches on a dataset (2023). [11] This strategy separates the real instances from the fake ones, keeps the data safe and stable, and provides machine learning models a lot of space to work and get better results. The fuzzy C-means under sampling approach was used to reduce data imbalance and improve the performance of fraud detection systems by keeping data variety high.

Received: July 21, 2025

Yilmaz A. A. (2023) [5] employs a machine learning approach integrated with Particle Swarm Optimization (PSO) to use pertinent information in credit card fraud detection. There are four basic steps: normalizing data, utilizing SMOTE to fix distorted data, PSO to choose the best features, and classification using decision tree, random forest, logistic regression, artificial neural networks, and naive Bayes method. The experimental findings of applying the suggested system to a European credit card dataset validated its exceptional accuracy (reaching 99.92%) and demonstrated its superiority over contemporary models across the majority of assessment criteria.

Du and others AE-XGB-SMOTE-CGAN 12: Suggested a way to fix the issue of class imbalance in finding credit card fraud In this model, auto encoders help find important features, XGBoost sorts them, and a mix of SMOTE and CGAN produces more examples of fraud. Our suggested model was able to enhance the accuracy by 2% and the MCC by 30% compared to KNN and LightGBM. It is likely that significant enhancements may not occur owing to the inherent complexity of the model and its concentration on a single dataset of financial transactions. But SMOTE's synthetic data after CGAN optimization may not have all the same traits as real fraudulent conduct.

The authors of (2024) [13] put out a novel algorithm for detecting credit card fraud using their Brown Bear Optimization (BBO) methodology. We created a binary BBOA and then used it with SVM, k-NN, and XGB Tree to test it on the Australian credit card dataset. Results of the Experiment The experimental findings indicate that our solution achieves 91% classification accuracy, reduces feature size by 67%, and surpasses 10 well evaluated benchmarking methods in the majority of performance parameters..

### 1.1 K-Means Clustering Algorithm

The K-Means algorithm is used to group data points into clusters by reducing the total distance to each cluster's centre which makes the data structure and analysis easier [16]. This algorithm divides a dataset into k groups, where each observation goes into the cluster closest to its centroid [17]. K-Means a mathematical problem, one needs to do the following:

- Objective Function (Inertia): The purpose of K-Means is to minimize how far apart data points and their cluster centroids are. The equation for the objective function is shown in equation 1.

$$J = \sum_{i=1}^{k} \sum_{x \in C_i} \|x = \mu_i\|^2 \tag{1}$$

Where: k is the number of clusters, Ci is the set of data points in the i-th cluster, μi is the centroid of the i-th cluster, x is a data point, ∥x−μi∥2 is the squared Euclidean distance between x and μi.

- Centroid Calculation: the centroid μi of the i-th cluster is computed as the mean of all data points assigned to that cluster, equation 2 show the equation of calculated centroid.

$$\mu_i = \frac{1}{|C_i|} \sum_{x \in C_i} x \tag{2}$$

Where: |Ci | is the number of data points in cluster Ci.

The steps of k-means algorithm method is:

Received: July 21, 2025

- step1: Define the k number of cluster.
- step2: Random initialize the centres.
- step3: Assign the point to nearest centre using equation 1.
- step4: recalculate the centres using equation 2.
- steps5: return to step3 until the number of iteration is reached or no change in centres.

### 1.2 Particle Swarm Optimization (PSO)

PSO is a swarm intelligence method that uses efficient exploration of the solution space to maximize cluster centres of gravity and increase clustering accuracy. This method is inspired by the social behaviour of birds. The PSO algorithm initialization parameters include the number of particles, maximum iterations, inertia weight (w), cognitive coefficient (c1), and social coefficient (c2). The velocities of all particles are set to zero, and each particle points to a possible solution (the cluster centre of gravity locations). To minimize the intra-cluster distance, the sum of the squares of the distances between data points and their specified centres of gravity, the algorithm iteratively changes the particle locations during the PSO optimization loop. The equation for updating the velocity of each particle provides in Equation 3 [5] [6]

$$v_p(t+1) = w \times v_p\,(t) + c_1 \times r_1 \times \left(perosnalBest_p - x_p(t)\right) + c_2 \times r_2 \times \left(globalBest - x_p(t)\right) \tag{3}$$

Where: vp(t) is the velocity of particle p at iteration t, xp(t) is its position, r1 and r2, are random factors, personalBestp and globalBest represent the best positions found by the particle and the swarm, respectively. The particle's position is updated as shown in equation 4.

$$x_p(t+1) = x_p(t) + v_p(t+1) \tag{4}$$

### 1.3 the Interquartile Range method (IQR)

The interquartile range method is used to identify and eliminate outliers, which may negatively impact model performance. The following are the mathematical instructions for finding and eliminating outliers using the interquartile range rule [15].

- Calculating Quartiles: For a given characteristic, arrange the data in ascending order and calculate the first quartile (Q1) and third quartile (Q3).

Q1: The 25th percentile of the data, or the value 25% of the data fall below it.
Q3: The 75th percentile of the data, or the value 75% of the data fall below it.

- Calculating the Interquartile Range (IQR): The interquartile range is the range between the first quartile (Q1) and the third quartile (Q3) is shown in equation 5.

$$IQR = Q3 - Q1 \tag{5}$$

- Determine the Lower and Upper Bounds: The lower and upper bounds are calculated to identify outliers as shown in equation 6 and 7 respectively. Any data point outside these bounds is considered an outlier.

Received: July 21, 2025

$$Q1 - 1.5 \times IQR \tag{6}$$

$$Q3 + 1.5 \times IQR \tag{7}$$

- Identifying Outliers: A data point x is considered an outlier if it meets one of the following two conditions:

$$x < Q1 - 1.5 \times IQR$$

$$x > Q3 + 1.5 \times IQR$$

- Remove Outliers: rows in the dataset where the feature value is an outlier are removed to clean the data.

### 1.4 Rescaling Data

The RobustScaler is a method used to standardize features by removing the median and scaling data by the IQR [15]. MinMaxScaler stays strong in the presence of outliers, unlike some other methods used for scaling, equation 8 is shown RobustScaler.

$$Xscaled = \frac{X - median(X)}{IQR(X)} \tag{8}$$

Where: median(X) is the feature X's median, IQR(X) is (Q3−Q1) the interquartile range of X, where Q1 is the 25th percentile of X and Q3 is the 75th percentile of X.

## 2.0 Methods And Materials

Figure 1 illustrates the proposed Hybrid method of PSO + K-means method. At the start, the analysis requires importing a dataset of credit card transactions, which includes information about amounts transacted, the time of the transactions, and relevant details. Usually, the raw data has too many unorganized variables, so it's important to select which features to use. It allows you to retain the variables that matter most for accurate detection of anomalies.

The proposed hybrid approach, Hyper Particle Swarm Optimization (PSO) and K-Means are integrated to determine optimal initial cluster centres for improved clustering accuracy and stability. Because the centroids are randomly assigned at the start, Traditional K-Means tends to have low levels of convergence and ends up with ineffective grouping. To solve this issue, PSO is employed to survey the entire space and find areas with good chances to have the best solutions. Using its global search feature, PSO places the first cluster centres in areas determined by the organization of the data set.

More of a "strangle a gem" approach to innovation: Particle-swarm optimization (PSO) is a one-two preserver that combines wide global seeding with fine-tuning local crystals. It is used with K-means clustering. So, pre-calibrated centroids via PSO are useful seeds for K-means round, which pulls related records into tightly knit clusters. First, this two-step process makes a significant difference in the problem of most local minima locking the analysis in shallow pits. The hybrid technique is now a good candidate for finding the outlier signals hidden deep in the high-dimensional feature spaces of credit card logs since the convergence behavior has improved.

Received: July 21, 2025

Any evident clusters that come out during analysis are quickly checked for outlier data, which gives them a second level of evaluation. K-Means makes reasonable clusters, however it doesn't give points that are far away from the centroid a name. To make up for it, the Interquartile Range is used to let you do a boxplot variability check on each condensed subset. Any transaction that ends up beyond the IQR fence is marked as one that has to be followed up on by brand compliance..

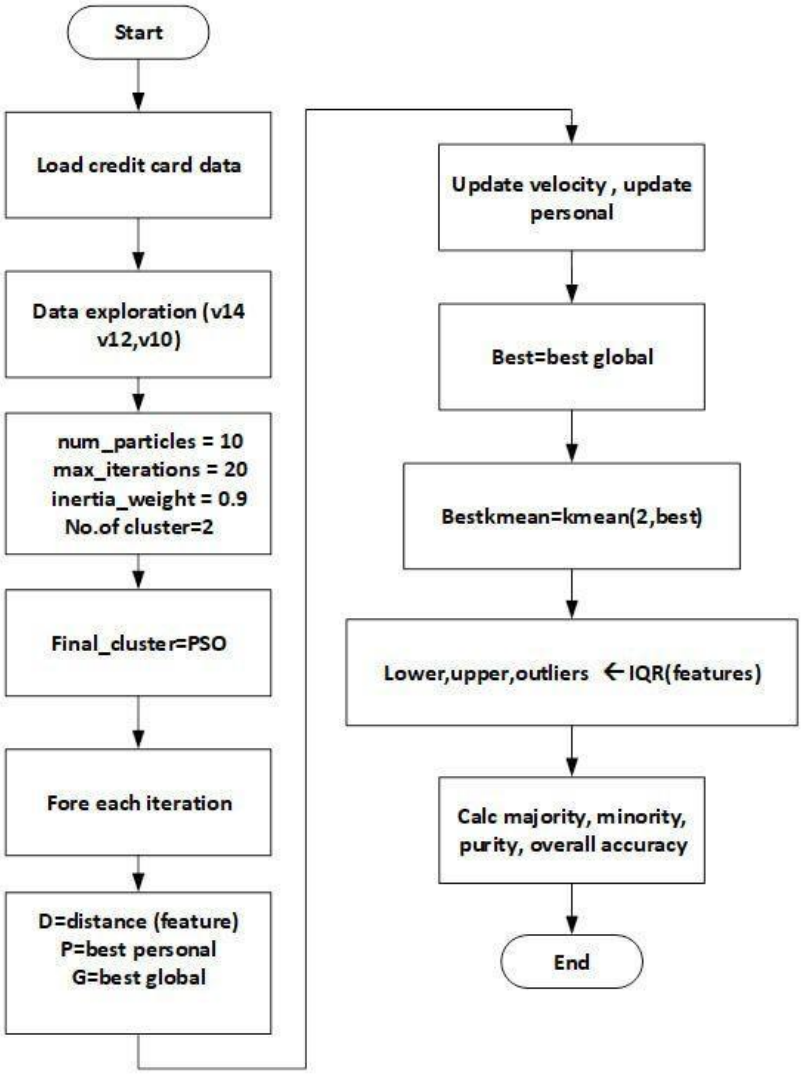


Figure 1. Proposed PSOKClus method

## 3.0 Results And Discussion

This part gives a summary of the dataset used, lists the evaluation measures, and shows the results of the proposed PSOKClus algorithm for detecting credit card fraud. Both qualitative and quantitative metrics assess the clustering coherence and the overall effectiveness of the strategy..

### 3.1 Dataset Description

Received: July 21, 2025

The assessment is based on the publicly available CreditCard.csv dataset, which is often used in research on payment trends and anomaly detection. The rows in this table are what make up the clusters and outlier points. Details for each column follow:

- Time: A millisecond-resolution stamp marking when money changed hands. Because the goal is pattern discovery rather than chronological mapping, this field is excluded from the clustering input.
- Variables (V1 through V28): Twenty-eight anonymized floating-point measures computed via PCA, each encoding a distinct transactional characteristic without revealing cardholder identity.
- Amount: The price-tag attached to the purchase, factored into irregularity scoring alongside the PCA dimensions.
- Category (Target Variable): A binary node that flags the row as fraudulent (1) or legitimate (0), serving as the ground truth for supervised evaluations.

The data was first pruned and scaled so every variable occupies a roughly comparable numerical range; this keeps distance computations straightforward and blocks any single feature from overpowering the others. Similar scaling is a standard safeguard against radically uneven value spreads [14][18].

### 3.2 Evaluation Metrics

Clustering quality is measured using several standard gauges., the following metrics are used:

**A. Silhouette Score:** observes how tightly each point clusters with its neighbors relative to the next closest assembly. The index glides from -1 to +1:

- Values near +1 suggest snug membership
- Numbers clustering around 0 imply a doorstep position between groupings.
- Negative readings hint that the item may have landed in the wrong collection altogether.

Mathematical Formulation:

- Intra-cluster distance:

$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i j \neq i} d(i,j) \tag{9}$$

Where: Ci is the cluster to which point i belongs, d(i,j) is the distance among points i and j. |Ci| is the number of points in cluster Ci

- Nearest-cluster distance:

$$b(i) = \min_{k \neq i} \left( \frac{1}{|C_k|} \sum_{j \in C_k} d(i,j) \right) \tag{10}$$

Where: $C_k$ is any cluster other than Ci.

- Silhouette Score:

$$s(i) \frac{b(i) - a(i)}{\max\,(a(i), b(i))} \tag{11}$$

Received: July 21, 2025

The overall Silhouette Score for the dataset is the average of s(i) for all points:

$$S = \frac{1}{N}\sum_{i=1}^{N} s(i) \tag{12}$$

Where: N is the total number of points.

**B. Davies-Bouldin Index (DBI):** This measure compares intra-cluster scatter and inter-cluster separation to assess the quality of clustering. Better clustering is indicated by lower values. Similar to Formulation

- Intra-cluster scatter:

$$S_i = \frac{1}{|C_i|}\sum_{j \in C_i} ||X_i - C_i|| \tag{13}$$

Where: xj is a data point in cluster Ci, ci is the centroid of cluster Ci, ‖·‖ is the Euclidean distance.

- Inter-cluster distance:

$$d_{i,j} = |c_i - c_j| \tag{14}$$

- Similarity ratio:

$$R_{i,j} = \frac{S_i - S_j}{d_{i,j}} \tag{15}$$

- Davies-Bouldin Index:

$$DBI = \frac{1}{K}\sum_{i=1}^{K} \max_{j \neq i} R_{ij} \tag{16}$$

Where: K is the number of clusters.

**C. Dunn Index**

When comparing the minimum inter-cluster distance to the maximum intra-cluster diameter, the Dunn Index evaluates the ratio. Greater values signify improved grouping. As Formulation:

- Inter-cluster distance:

$$d_{min} = \min_{i \neq j} d(C_i, C_j) \tag{17}$$

Where: d(Ci,Cj) is the distance among clusters Ci and Cj , often defined as:

$$d(C_i, C_j) = \min_{x \in C_i, y \in C_j} |x - y| \tag{18}$$

- Intra-cluster diameter:

$$d_{max} = \max_{1 \leq i \leq K} diam\ (C_i) \tag{19}$$

Received: July 21, 2025

Where (diam (Ci)) is the diameter of cluster Ci, defined as:

$$diam\ (C_i) = \max_{x,y \in C_i} |x - y| \tag{20}$$

- Dunn Index:

$$Dunn\ Index = \frac{d_{min}}{d_{max}} \tag{21}$$

## 3.3 Results

In this section, there are two parts of comparison result: firstly evaluate the performance of clustering, secondly evaluate the fraud detection.

### 4.3.1 Evaluate the Performance of Clustering

In Table 1 presents the comparison result of proposed PSOKClus and K-means methods using three metrics: Silhouette Score, Davis-Bolden index and Dunn index.

Table 1 shows the highlight score of Silhouette Score (0.8552) of proposed PSOKClus method which indicates strong within-cluster cohesion and good cluster separation, demonstrating good clustering quality compare than k-means which is (0.8512). The metrics of Davis-Bolden index, which scored (0.4770) in proposed PSOKClus is lower than k-means (0.5683) that is confirmed this by showing low within-cluster dispersion and high between-cluster distance. The Dunn index value of proposed PSOKClus is (0.0238) is higher than k-means (0.0147) indicates a slight difference in cluster separation.

Table 1. Comparison result of proposed PSOKClus and K-means methods

| Metrics | PSOKClus | K-means |
|---|---|---|
| Silhouette Score | **0.8552** | 0.8512 |
| Davies-Bouldin Index | **0.4770** | 0.5683 |
| Dunn Index | **0.0238** | 0.0147 |

### 4.3.2 Evaluate the Fraud Detection

In Tables (2) PSOKClus detected significantly fewer outliers in cluster (0) (V14:3, V12:0, V10:7) than in same cluster with kmeans (V14: 14032, V12: 15278, V10:9051) as represented in table (4). Also, PSOKClus of cluster (1) (V14: 213, V12: 161, V10: 293) as represented in table (3) detect significantly than K-Means of same cluster (1) (V14: 183, V12:133, V10:162) as represents in tables (5). This indicates that the enhanced PSOKClus points form compact and homogeneous clusters. The number of outliers (the number of outliers/total cluster size) is 0.05% for the PSO-KMeans 1 cluster compared to 5% for the K-Means 0 cluster, indicating that normal clustering is better.

Received: July 21, 2025

Table 2. Outlier Detection for Cluster 0 using PSOKClus method

| Metrics | V14 Outlier Detection | V12 Outlier Detection | V10 Outlier Detection |
|---|---|---|---|
| **Quartile 25** | -13.2542 | -14.1698 | -13.1003 |
| **Quartile 75** | -8.9889 | -8.6775 | -7.1875 |
| **IQR -** | 4.2652 | 5.4924 | 5.9133 |
| **Cut Off** | 6.3978 | 8.2386 | 8.8700 |
| **Lower bound** | -19.6520 | -22.4084 | -21.9708 |
| **Upper bound** | -2.5911 | -0.4389 | 1.6824 |
| **Number of outliers** | 3 | 0 | 7 |

Table 3. Outlier Detection for Cluster 1 using PSOKClus method

| Metrics | V14 Outlier Detection | V12 Outlier Detection | V10 Outlier Detection |
|---|---|---|---|
| **Quartile 25** | -4 .1397 | -2.0167 | -1.9989 |
| **Quartile 75** | -2.9524 | -0.3755 | -1.0490 |
| **IQR -** | 1.1873 | 1.6412 | 0.9500 |
| **Cut Off** | 1.7809 | 2.4618 | 1.4250 |
| **Lower bound** | -5.9206 | -4.4785 | -3.4239 |
| **Upper bound** | -1.1715 | 2.0863 | 0.3760 |
| **Number of outliers** | 213 | 161 | 293 |

Table 4. Outlier Detection for Cluster 0 using K-means method

| Metrics | V14 Outlier Detection | V12 Outlier Detection | V10 Outlier Detection |
|---|---|---|---|
| **Quartile 25** | -0.4285 | -0.4018 | -0.5252 |
| **Quartile 75** | 0.4875 | 0.6187 | 0.4604 |
| **IQR -** | 0.9160 | 1.0206 | 0.9856 |
| **Cut Off** | 1.3740 | 1.5308 | 1.4784 |
| **Lower bound** | -1.8025 | -1.9327 | -2.0036 |
| **Upper bound** | 1.8615 | 2.1496 | 1.9387 |
| **Number of outliers** | 14032 | 15278 | 9051 |

Received: July 21, 2025

Table 5. Outlier Detection for Cluster 1 using K-means method

| Metrics | **V14 Outlier Detection** | **V12 Outlier Detection** | **V10 Outlier Detection** |
|---|---|---|---|
| **Quartile 25** | -0.2126 | -0.5885 | -1.2351 |
| **Quartile 75** | 0.7621 | 0.5862 | 0.1756 |
| **IQR -** | 0.9747 | 1.1747 | 1.4107 |
| **Cut Off** | 1.4621 | 1.7620 | 2.1161 |
| **Lower bound** | -1.6747 | -2.3504 | -3.3511 |
| **Upper bound** | 2.2241 | 2.3482 | 2.2917 |
| **Number of outliers** | 183 | 133 | 162 |

The results in Table 6 indicate that the proposed method, PSOKClus (a combination of PSO and K-Means), delivered better performance in classifying fraudulent transactions compared to the traditional K-Means method. When comparing the distribution of classes within the clusters, we observe that PSOKClus grouped 154 fraud cases in cluster 0 and 52 cases in cluster 1, totaling 206 detected fraud cases out of 492, within a dataset of 1,486 transactions (184 in cluster 0 and 1302 in cluster 1).

In contrast, the K-Means method was only able to cluster 43 fraud cases (40 in cluster 0 and 3 in cluster 1) out of 492, within a much larger dataset of 252,930 transactions (248,090 in cluster 0 and 4,840 in cluster 1).

Moreover, PSOKClus achieved high cluster purity rates of 99.94% and 99.98% in its two clusters, compared to K-Means, which had purity rates of 99.84% and 96.94%. This 2% to 3% difference in purity is considered significant in fraud detection applications due to the sensitivity of results and the risks posed by false positives.

Additionally, the higher overall alignment accuracy achieved by PSOKClus—99.98% compared to 98.072% for K-Means—demonstrates the reliability of the hybrid method in classifying data, especially in an imbalanced data environment.

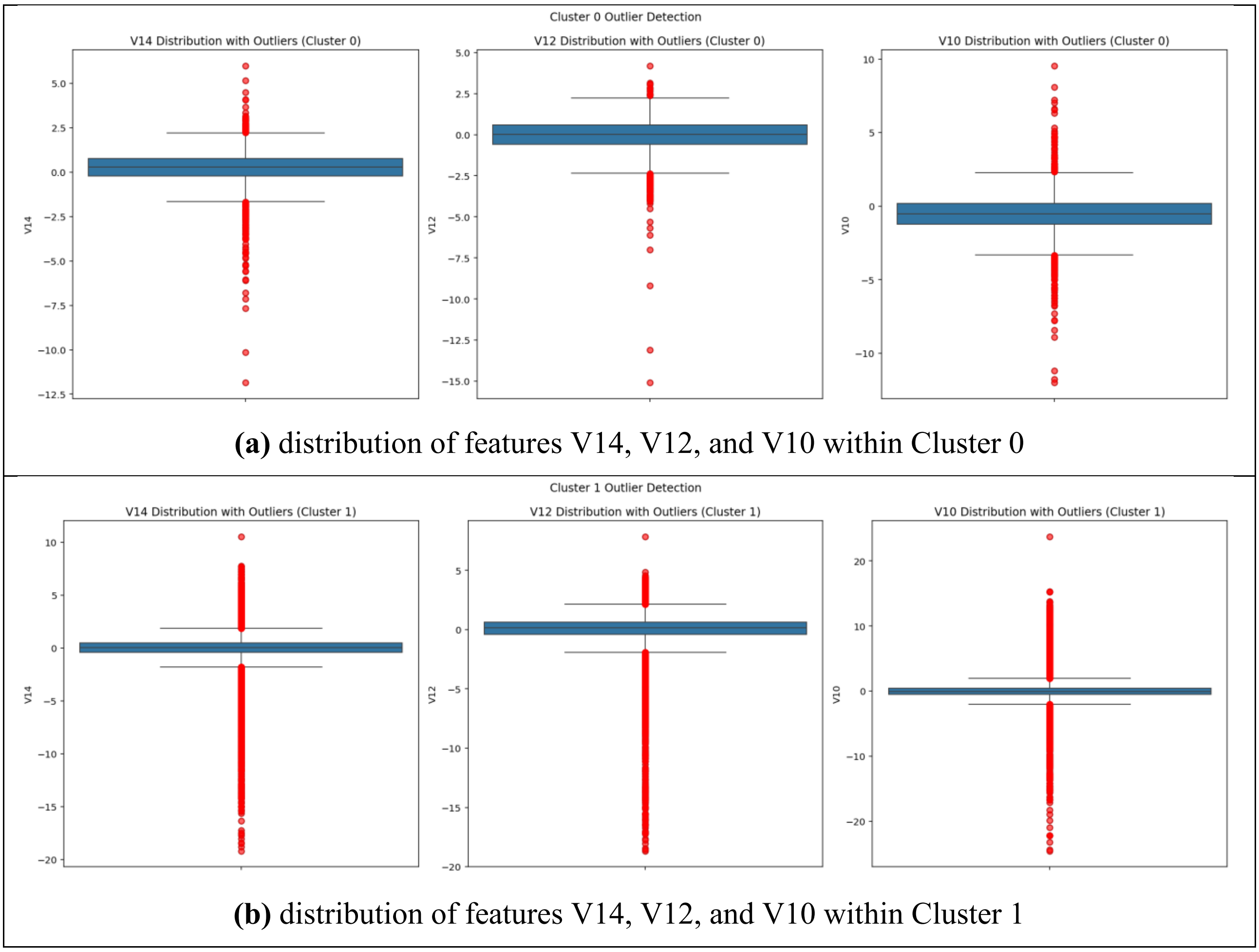
Table 6. Cluster-to-Class Distribution Comparison result of proposed PSOKClus and K-means methods

| Methods | Cluster | Total Samples | Class 0 | Class 1 | Majority | Purity |
|---|---|---|---|---|---|---|
| proposed PSOKClus | 0 | 184 | 30 | 154 | 1 | 0.999381 |
| | 1 | 1302 | 1250 | 52 | 0 | 0.999839 |
| Overall Alignment Accuracy of proposed PSOKClus: 99.98% | | | | | | |
| K-means | 0 | 248090 | 248050 | 40 | 0 | 0.99839 |
| | 1 | 4840 | 4837 | 3 | 0 | 0.969380 |

Received: July 21, 2025

Overall Alignment Accuracy of K-means: 98.072%

Figure 2 shows the Cluster Analysis with Outlier Awareness in proposed PSOKClus. As shown in Figure (2.a) illustrates the distribution of features V14, V12, and V10 within Cluster 0, highlighting the presence of outliers based on the IQR method. Each boxplot displays the central 50% of the data (IQR), with whiskers indicating 1.5 times the IQR from the quartiles. Data points beyond this range are marked in red and considered outliers. Notably, Cluster 0 shows a relatively small number of outliers, suggesting a more homogeneous structure and compact distribution, which aligns with its lower variance and higher purity as indicated in the clustering analysis.



**(a)** distribution of features V14, V12, and V10 within Cluster 0

**(b)** distribution of features V14, V12, and V10 within Cluster 1

Figure 2. Outlier Detection across Clusters Using IQR on Key Features (V10, V12, V14)

Figure (2.b) presents the same outlier analysis for Cluster 1, focusing on features V14, V12, and V10. Compared to Cluster 0, the number of outliers is significantly higher across all three features, especially in the negative range. The dense spread of red points indicates greater variability within this cluster, which may be attributed to the large volume of transactions it contains. These outliers could represent anomalous but legitimate behaviour or potential fraudulent patterns requiring further investigation. The figure reinforces the importance of outlier filtering in improving the reliability of the clustering-based fraud detection model.

Figure 3 presents a visual summary of the clustering results. The left panel shows the PCA-projected clusters with outliers marked in red, highlighting points that deviate significantly from normal patterns—primarily around Cluster 1, which contains most of the data. The right panel illustrates the original class distribution per cluster, revealing a strong alignment between the unsupervised clusters and actual class labels. This indicates that the PSOKClus method successfully identifies distinct transaction behaviors and isolates anomalies, making it highly suitable for real-world fraud detection applications.

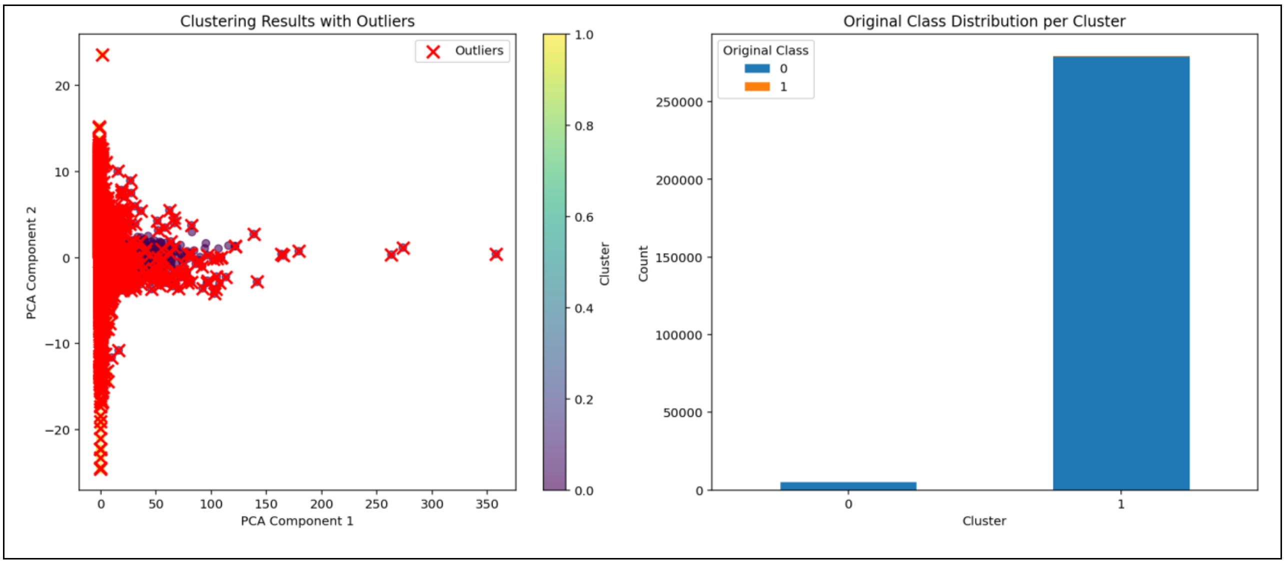


Figure 3. Clustering Results and Class Distribution with Outliers

## 4.0 Conclusion

In this study, a hybrid clustering model combining PSO with K-Means to detect anomalies in credit card transactions is proposed. By leveraging the optimization capability of PSO for cluster center initialization, the model achieved significantly improved clustering performance, as demonstrated by high Silhouette Scores and low Davies-Bouldin Index values. Our study using real data found that the PSOKClus method successfully separates normal transactions from fraudulent transactions, with an alignment accuracy of 99.98%. Besides, including an IQR-based check helped the model pick up small anomalies that regular methods might overlook. The tests prove that combining meta-heuristic optimization with traditional clustering creates a reliable, scalable solution for fast fraud detection. This way of working helps spot minority cases in datasets with few examples and it lowers the risk of false positives—a major requirement for financial use cases. Going forward, we plan to integrate analysis of time-based changes and instant deployment to improve both our detection accuracy and operations. In addition, PSOKClus shows good computational efficiency and achieves almost perfect cluster purity, significantly reducing false positives—a key factor in reducing financial losses. With only 0.016% contamination and low number of false positives, PSOKClus method is an effective and efficient clustering method in fraud detection system. In the future, DL might be used to evaluate streaming data to make fraud detection systems work better and faster.  Using both dynamic thresholding and hybrid ensemble methods may also assist reduce false positives and improve detection accuracy in datasets that aren't evenly distributed.

Received: July 21, 2025

## Conflict Of Interest

The authors declare no conflicts of interest.

## References

[1] Du, HaiChao, Li Lv, Hongliang Wang, and An Guo., "A novel method for detecting credit card fraud problems," *PloS one 19,* no. e0294537, p. 3, 2024.

[2] H Palivela, V Rishiwal, S Bhushan, A Alotaibi, "Optimisation of Deep Learning based Model for Identification of Credit Card Frauds," *ieeexplore,* vol. 12, no. 1, pp. 25629 - 125642, 2024.

[3] Zhu M, Zhang Y, Gong Y, Xu C, Xiang Y., "Enhancing credit card fraud detection a neural network and smote integrated approach.," *arXiv preprint arXiv:2405.00026.,* 2024 Feb 27.

[4] R., PK, "Enhanced Credit Card Fraud Detection: A Novel Approach Integrating Bayesian Optimized Random Forest Classifier with Advanced Feature Analysis and Real-time Data Adaptation.," *International Journal for Innovative Engineering & Management Research, Forthcoming,* 2023 May 28..

[5] Yılmaz AA. , "A machine learning-based framework using the particle swarm optimization algorithm for credit card fraud detection.," *Communications Faculty of Sciences University of Ankara Series A2-A3 Physical Sciences and Engineering,* vol. 66, no. 1, pp. 82-94, 2023.

[6] Kuncoro, M. W., & Suharjito, "Boruta feature selection and particle swarm optimization," *Journal of Theoretical and Applied Information Technology,* vol. 100, no. 8, pp. 2578-2585, 2022.

[7] H Zhou, M Zhang, L Pang, JH Li, "Abnormal detection of cash-out groups in IoT based payment," *mdpi,* vol. 4, no. 3, pp. 1-16, 2021.

[8] VT Manda, D Kondapalli, A sai Malla, NM Jyothi, "Imbalanced Data Challenges and Their Resolution to Improve Fraud Detection in Credit Card Transactions," *researchsquare,* vol. 2, no. 1, pp. 1-13, 2024.

[9] Mniai A, Jebari K. , "Credit card fraud detection by improved SVDD.," *InProceedings*

Received: July 21, 2025

*of the world congress on engineering,* pp. 6-8, 2022 .

[10] N Prabhakaran, R Nedunchelian, "Oppositional Cat Swarm Optimization-Based Feature Selection Approach for Credit Card Fraud Detection," *Wiley Online Library,* vol. 19, no. 10, pp. 1-13, 2023.

[11] H Ahmad, B Kasasbeh, B Aldabaybah, "Class balancing framework for credit card fraud detection based on clustering and similarity-based selection (SBS)," *Springer,* vol. 15, no. 1, p. 325–333, 2023.

[12] HC Du, L Lv, H Wang, A Guo, "A novel method for detecting credit card fraud problems," *Plos one,* vol. 19, no. 3, pp. 1-26, 2024.

[13] Sorour SE, AlBarrak KM, Abohany AA, Abd El-Mageed AA., " Credit card fraud detection using the brown bear optimization algorithm," *Alexandria Engineering Journal,* vol. 104, pp. 171-92., 2024 Oct 1.

[14] Yong Fang, Yunyun Zhang, Cheng Huang, "Credit Card Fraud Detection Based on Machine Learning," *Computers, Materials & Continua,* vol. 61, no. 1, pp. 185-195, 2019.

[15] Yan C, Wang J, Zou Y, Weng Y, Zhao Y, Li Z., " Enhancing credit card fraud detection through adaptive model optimization.," *In2024 IEEE 7th International Conference on Big Data and Artificial Intelligence (BDAI),IEEE.,* pp. 49-54, 2024 Jul 5.

[16] MS Yang, I Hussain , "Unsupervised multi-view K-means clustering algorithm," *IEEE Access,* vol. 11, pp. 13574-13593, 2023.

[17] Sarkar M, Puja AR, Chowdhury FR., " Optimizing marketing strategies with RFM method and K-means clustering-based AI customer segmentation analysis.," *Journal of Business and Management Studies,* vol. 6, no. 2, pp. 54-60, 2024 Mar 7.

[18] https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud

Received: July 21, 2025