IJDACR
ISSN: 2319-4863

IJDACR
International Journal Of Digital Application & Contemporary Research

**International Journal of Digital Application & Contemporary Research**
**Website: www.ijdacr.com (Volume 10, Issue 07, February 2022)**

# Credit Card Fraud Detection using BPSO based Features and Random Forest Classifier

Dr. Hemant N. Patel
Assistant Professor,
Computer Engineering,
Sankalchand Patel College of Engineering,
hp15284@gmail.com

Dr. Amit N. Patel
Associate Professor,
MBA Department,
Sankalchand Patel College of Engineering
amitnpatel1@gmail.com

Mr. Sunil P. Patel
Assistant Professor,
Sankalchand Patel College of Engineering,
sunilpatel148@gmail.com

*Abstract* - **Credit card fraud is a social problem that faces many ethical challenges and poses a serious threat to businesses around the world. Machine learning algorithms are used to detect fraudulent transactions by authors. This study presents an implementation of an automated credit card fraud detection system where pre-processing the data, among others. Binary particle swarm optimization (BPSO) algorithm are used for selection of features with random forest (RF) classifier for training and testing from Kaggle dataset. Sensitivity, precision, f-score and accuracy are used as a performance evaluation tool to evaluate the proposed technique.**

*Keyword* - **RF, Kaggle, BPSO, etc.**

## I. INTRODUCTION

Cards, be they credit or debit cards, are means of payment highly used in commerce in general. Even so, there is room for them to be used even more. These factors arouse the interest of fraudsters. In a first analysis, when a fraud is successful, it could be said that the companies involved bear the costs, but, in fact, this operational loss makes prices more expensive for the final consumer. Hence, combating fraud in this payment method brings benefits to the whole society. This task, due to its complexity and size, has to be done in a concomitantly effective and efficient way. Therefore, there is an opportunity to apply machine learning techniques.

We use this method as our primary resource development method for detecting credit card fraud. However, this resource engineering strategy raises several research questions. We initially assumed that these descriptive statistics could not fully describe the sequential nature of fraud and authentic patterns, and that simulating the sequence of transactions might be useful for detecting fraud. In addition, expert knowledge is used to create these additional capabilities, and sequence modeling can be automated with class labels available for previous transactions. Finally, the aggregate characteristics are point estimates that can be complemented by a multidisciplinary description of the transaction context (especially from the merchant's point of view).

## II. LITERATURE REVIEW

For fraud detection, a variety of supervised and semi-supervised machine learning techniques are used [8], but our goal is to overcome three main challenges with the card fraud dataset, namely, strong class imbalance, the inclusion of labelled and unlabelled samples, and the ability to process a large number of transactions.

Decision Trees, Naive Bayes Classification, and Least Squares Classification are examples of supervised machine learning algorithms [3].

In real-time datasets, Squares Regression, Logistic Regression, and SVM are utilized to detect fraudulent transactions.

To train the behavioral aspects of normal and aberrant transactions, two approaches under random forests [6] are utilized.

Random-tree-based random forest and CART-based random forest are the two types. Despite the fact that random forest produces good results. In the case of imbalanced data, even with a tiny set of data, there are still certain issues. The focus of future effort will be on resolving the problem.

On severely skewed data, the performance of Logistic Regression, K-Nearest Neighbor, and Nave Bayes is investigated, where research is being done on investigating meta-classifiers and meta-learning techniques in credit card fraud data dealing with credit card fraud data that is severely unbalanced.

IJDACR
ISSN: 2319-4863

IJDACR
International Journal Of Digital Application & Contemporary Research

**International Journal of Digital Application & Contemporary Research**
**Website: www.ijdacr.com (Volume 10, Issue 07, February 2022)**

## III. PROPOSED METHOD

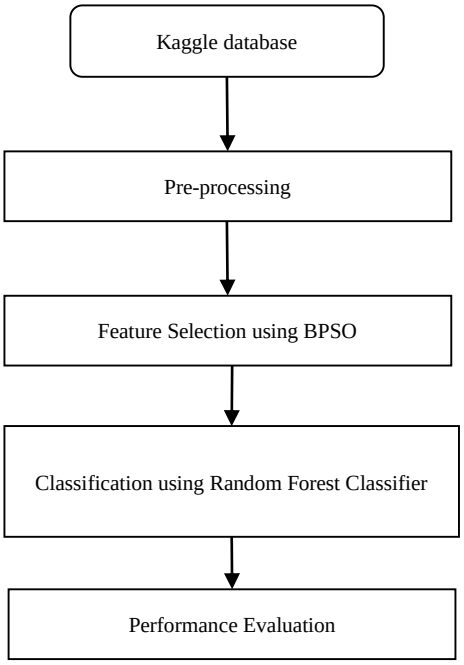


Figure 1: Flow diagram of proposed research work

***Data Collection and Data Analysis***

Kaggel data collection includes credit card transactions made in September 2013 by European cardholders. This database shows actions completed in two days and 492 out of 284,807 transactions. Inaccurate database, positive type (frauds) accounts for 0.172% of all transactions (Credit Card Fraud Detection Dataset).

In this research work, publically available benchmark datasets are used for credit card fraud detection

https://www.kaggle.com/mlg-ulb/creditcardfraud

Kaggle data collection includes credit card transactions made in by European cardholders. This database shows actions completed in two days and 492 out of 284,807 transactions. Inaccurate database, positive type (frauds) accounts for 0.172% of all transactions.

***Pre-Processing***

- Create a categorical response variable where:
  - 1= fraud activity
  - 0 = normal or non-fraudulent activity
- Time attribute removal from the categorical class

***Feature Selection using Binary Particle Swarm Optimization***

Features selection can be applied to enable data analysis, in addition to providing the most accurate classification. There are several reasons for reducing the features, such as: (1) avoiding overfitting and improving the model's performance; (2) providing faster models and using less resources, (3) providing a better understanding of the processes that generated the data. Therefore, a strategy is to select the most relevant features of the studied context. Here the above-mentioned selection of features is accomplished by binary version of particle swarm optimization i.e. BPSO. The PSO is an intuitive method. It is a complex algorithm that involves certain herds in nature. A particle swarm optimization algorithm created by observing the behavior of populations of birds, fish and bees (Kennedy and Eberhart, 1997).

Binary particle swarm optimization has been introduced by (Kennedy and Eberhart, 1997). If the elements of the problem can be sorted or grouped, binary swarm optimization can be used to solve this discrete problem (Khanesar et al., 2007). Investigations of many optimization problems, such as routing or scheduling problems, are carried out in discrete spaces.

For the search area $S = \{0,1\}^D$, the fitness function $f$ maximizes i.e. $(\max f(x))$. The $i^{th}$ particle in $D$ dimension is defined as:

$$X_i = (x_{i1}, x_{i2}, \dots x_{id})^T,\ x_{id} \in \{0,1\}, d = 1,2, \dots, D \quad (1)$$

The velocity vector in $D$ dimension can be represented as:

$$V_i = (v_{i1}, v_{i2}, \dots v_{id})^T,\ v_{id} \in [-V_{max}, V_{max}], d = 1,2, \dots, D \quad (2)$$

Where $V_{max}$ is the vector of the maximum velocity.

The best position at the moment can be characterized as (Khanesar et al., 2007):

$$p_i = (p_{i1}, p_{i2}, \dots p_{id})^T,\ p_{id} \in \{0,1\}, d = 1,2, \dots, D \quad (3)$$

The definition according to the above notation can be given as follows:

Equation for Velocity:

$$v_{id} = v_{id} + c_1 rand_1(p_{id} + x_{id}) + c_2 rand_2(p_{gd} - x_{id}) \quad (4)$$

Equation for position:

$$X_{id} = \begin{cases} 1 & if\ U(0,1) < sigm(v) \\ 0 & otherwise \end{cases},\ d = 1,2, \dots D; i = 1,2, \dots, N \quad (5)$$

Transfer function:

$$sigm(v_{id}): \frac{1}{1+\exp(-\lambda v_{id})} \quad (6)$$

$g$: index of the best performing particle

$p_{gd}$: best part

$N$: the width of the fortification

IJDACR
ISSN: 2319-4863

IJDACR
International Journal Of Digital Application & Contemporary Research

International Journal of Digital Application & Contemporary Research
Website: www.ijdacr.com (Volume 10, Issue 07, February 2022)

$c_1, c_2$: social and cognitive component constants

$rand_1, rand_2$: $U(0,1)$ random numbers

$sigm(v_{id})$: sigmoid transform function

***Classification Algorithms***

**Random Forest Classifier.** In the random forest (RF) approach, a combination of several decision trees is made to make the predictions. The aim of this approach is to improve the learning process and consequently the overall result.

The RF algorithm works by randomly selecting $k$ elements from the total training set. In the next step, these $k$ elements are used to find the root node, which consists of the element that best separates the groups considered. In the next step, new elements are added in each node of the tree separation, until the leaf node is obtained, which consists of the classification result for that specific tree. This process is repeated, randomly creating n decision trees (Breiman, 2001).

In the test step, the sets are submitted to the rules created for each decision tree, obtaining results for each tree. An analysis of all the outputs obtained by the decision trees is performed and the final result consists of the output that was predicted by more trees. Following are the reasons to consider the random forest classifier for this research work:

- Random forest reduces overfitting in decision trees and helps improve accuracy.
- It is flexible in terms of classification and regression tasks.
- It goes well with categorical and persistent values and Automate missing values in the data.

A random forest is a classifier consisting of a set of base classifiers such as a decision tree shown:

$$\{h(x, \Theta_k), \quad k = 1, \dots L\}$$

Random forests are composed of a set of binary decision trees in which randomness has been introduced.

Random forests were introduced by (Breiman, 2001) by the following very general definition:

Let $\left(\hat{h}(\Theta_1), \dots, \hat{h}(\Theta_q)\right)$ a collection of tree predictors, with $\Theta_1, \dots, \Theta_q$ random variables independent of $\mathcal{L}_n$. The predictor of random forests $\hat{h}_{RF}$ is obtained is aggregating this collection of random trees as follows:

- $\hat{h}_{RF}(x) = \frac{1}{q}\sum_{l=1}^{q} \hat{h}(x, \Theta_l)$ (7)
  Average of individual tree predictions in regression.
- $\hat{h}_{RF}(x) = \arg\max_{1 \le k \le K} \sum_{l=1}^{q} 1_{\hat{h}(x,\Theta_l)=k}$ (8)
- Majority vote among individual predictions trees in classification.

The term random forest comes from the fact that individual predictors are, here, explicitly predictors per tree, and that each tree depends on an additional random variable (that is, in addition to $\mathcal{L}_n$).

Optimizing the parameters of a machine learning model is an important step towards accurate predictions. Parameters define the properties of the model. Bayesian optimization is much more effective at setting the parameters of a machine learning based algorithm than random search methods (Zhang et al., 2021).

## IV. SIMULATION RESULTS

The proposed credit card fraud detection approach is simulated using MATLAB 2019a.

**Evaluation Parameters.**

The confusion matrix, composed of the first four values: True positive, false negative, false positive and True negative. The matrix was very useful, mainly for two reasons: first because its data described the result of the classification of each credit card fraud, and second because it is through it that the other metrics were obtained.

Table 1: Evaluation parameters

| | |
|---|---|
| TP (True Positive) | "Indicated the number of credit card fraud that were classified as correctly classified" |
| TN (True Negative) | "Indicated the number of credit card fraud that were classified as not classified correctly" |
| FP (False Positive) | "Indicated the number of credit card fraud that were classified as incorrectly classified" |
| FN (False Negative) | "Indicated the number of credit card fraud that were classified as not classified incorrectly" |

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (9)$$

$$Precision = \frac{TP}{TP+FP} \quad (10)$$

$$Sensitivity = \frac{TP}{TP+FN} \quad (11)$$

$$F - Score = \frac{2TP}{2TP+FP+FN} \quad (12)$$

IJDACR
ISSN: 2319-4863

IJDACR
International Journal Of Digital Application & Contemporary Research

**International Journal of Digital Application & Contemporary Research**
**Website: www.ijdacr.com (Volume 10, Issue 07, February 2022)**

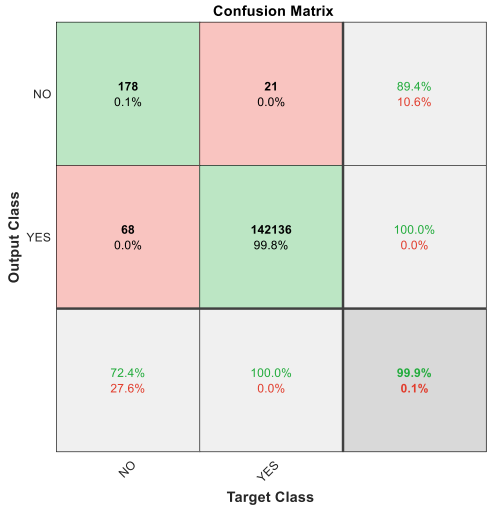


Figure 2: Confusion matrix plot for Bayesian optimized random forest classifier based credit card fraud detection

Here, TP=142136, TN=178, FP=68 and FN=21

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} = \frac{142136+178}{142136+178+68+21} = 99.93\%$$

$$Precision = \frac{TP}{TP+FP} = \frac{142136}{142136+68} = 99.95\%$$

$$Recall\ or\ Sensitivity = \frac{142136}{142136+21} = 99.98\%$$

$$F-Score = \frac{2TP}{2TP+FP+FN} = \frac{2\times142136}{2\times142136+68+21} = 99.96\%$$

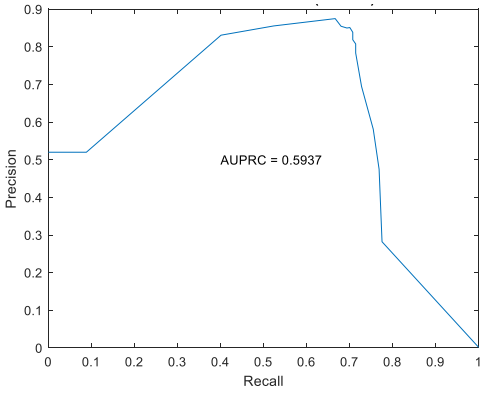


Figure 3: Precision-Recall curve

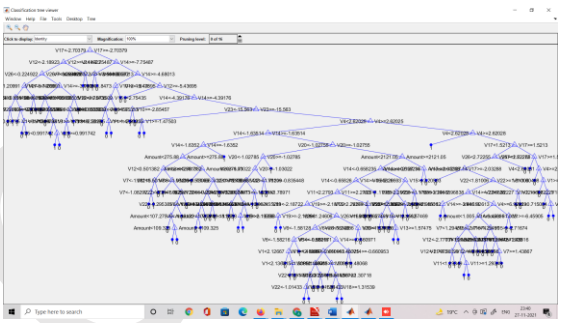

Figure 4: Random Forest tree

Table 2: Comparison with previous research works

| Method Used | | Accuracy | Precision | Sensitivity | F-Score |
|---|---|---|---|---|---|
| [9] | Logistic Regression | 90% | - | - | - |
| | Decision tree | 94.3% | - | - | - |
| | Random Forest Classifier | 95.5% | - | - | - |
| [8] | J48 Classifier | 74.6% | 76.5% | 91.9% | - |
| | AdaBoost M1 | 74.5% | 77.6% | 89.3% | - |
| | Random forest | 73.9% | 78.7% | 86% | - |
| | Naïve Bayes | 75.0% | 77.9% | 89.7% | - |
| Proposed Random Forest Classifier based approach | | 99.93% | 99.95% | 99.98% | 99.96% |

## V. CONCLUSION

The research for this paper is focused on the application of automatic learning techniques in the detection of credit card fraud: a difficult problem, motivated in the financial industry and highly applicable. There were abundant data available, but limited in terms of the number and diversity of variables. The investigation began with a critical analysis of the state of the art in fraud detection and the previous modeling procedure, which concluded with the proposal of several improvements. Then, the possibility of using approaches related to the detection of using supervised learning was explored to construct a useful score for the detection. Following this path, an important contribution of this work was produced: the development of a random forest classifier based credit card fraud detection. With the accuracy of 99.93%, it was found that the proposed approach is useful for fraud detection and to achieve good discrimination between classes.

References

[1] Sailusha, R., Gnaneswar, V., Ramesh, R., & Rao, G. R. (2020, May). Credit card fraud detection using machine learning. In *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)* (pp. 1264-1270). IEEE.

IJDACR
International Journal Of Digital Application & Contemporary Research

**International Journal of Digital Application & Contemporary Research**
**Website: www.ijdacr.com (Volume 10, Issue 07, February 2022)**

[2] Malini, N., & Pushpa, M. (2017, February). Analysis on credit card fraud identification techniques based on KNN and outlier detection. In *2017 third international conference on advances in electrical, electronics, information, communication and bio-informatics (AEEICB)* (pp. 255-258). IEEE.

[3] Kiran, S., Guru, J., Kumar, R., Kumar, N., Katariya, D., & Sharma, M. (2018). Credit card fraud detection using Naïve Bayes model based and KNN classifier. *International Journal of Advance Research, Ideas and Innovations in Technoloy*, *4*(3).

[4] Asha, R. B., & KR, S. K. (2021). Credit card fraud detection using artificial neural network. *Global Transitions Proceedings*, *2*(1), 35-41.

[5] Itoo, F., & Singh, S. (2021). Comparison and analysis of logistic regression, Naïve Bayes and KNN machine learning algorithms for credit card fraud detection. *International Journal of Information Technology*, *13*(4), 1503-1511.

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

[7] Nancy, A. M., Kumar, G. S., Veena, S., Vinoth, N. S., & Bandyopadhyay, M. (2020, November). Fraud detection in credit card transaction using hybrid model. In *AIP Conference Proceedings* (Vol. 2277, No. 1, p. 130010). AIP Publishing LLC.

[8] Singh, A. and Jain, A., 2019. Adaptive credit card fraud detection techniques based on feature selection method. In *Advances in computer communication and computational sciences* (pp. 167-178). Springer, Singapore

[9] Lakshmi, S.V.S.S. and Kavilla, S.D., 2018. Machine learning for credit card fraud detection system. *International Journal of Applied Engineering Research*, *13*(24 Pt. 1), pp.16819-16824.

IJDACR