# RELATIVE ANALYSIS OF RANDOM FOREST CLASSIFICATION OVER LINEAR REGRESSION CLASSIFIER TO DETECT CYBER THEFTS IN CREDIT CARD TO REDUCE FALSE RATE

**E Madhan Mohan[1], S. John Justin Thangaraj[2*]**

**Article History: Received:** 12.12.2022 **Revised:** 29.01.2023 **Accepted:** 15.03.2023

**Abstract**

**Aim:** To decrease the error rate of credit card cyber thefts based on binary selection of Novel Random Forest classifier and Linear Regression. The primary scope of this project is to find the malware attacks and it should be informed to the card holder and investigator simultaneously. **Materials and Methods:** Classification is performed by Random forest classifier(N=34) over Linear Regression (N=34) is for false rate detection. The statistical test difference between g-power 0.08 and alpha values are $\alpha$=0.05.
**Results:** The Independent sample T test is applied for the data set fixing confidence interval as 95%.
**Discussion and Conclusion:** Comparison has been made between the Novel Random forest classifier and the Linear Regression for this analysis. The accuracy of the random forest classifier is 94.4% and the Linear Regression accuracy of 51.9%.

**Keywords:** Novel Random Forest Classifier,Linear Regression, Machine learning, Fraud Detection, Credit card, Transaction.

[1]Research Scholar, Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamilnadu. India. Pincode: 602105.
[2*]Project Guide, Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamilnadu. India. Pincode: 602105.

## 1. Introduction

Fraud in credit card is an unapproved and undesirable use of some other account holder. An individual purpose somebody's card for individual requirements without knowing the proprietor of that specific account holder. Credit card fraud detection recognizes unlawful exchanges to get cash. Credit card fraud is one of the most advanced methods of money thefting ("Credit Card Fraud Detection System: A Survey" 2020; Padvekar et al. 2016). Owner and card issuing authorities are unaware that the card is being used by someone. Certain actions and controlling measures shall be taken in order to stop such type of fraudulent activities in the upcoming days. ("Credit Card Fraud Detection System: A Survey" 2020; Padvekar et al. 2016; Filippov, Mukhanov, and Shchukin 2008)). Fraud detection methods are continuously developed to defend the fraudulent activities from the unauthorized person activity ("Credit Card Fraud Detection System: A Survey" 2020; Padvekar et al. 2016; Filippov, Mukhanov, and Shchukin 2008; Akinbohun and Atanlogun 2018)). The applications of the credit card fraud detection system are finding implicit and hidden correlations in data, real time processing, the reduced number of verification measures, and detection of possible fraud scenarios. And to help the banks to perform the payments in a high range without any interruption.

Total number of articles published are 220 in the last 5years articles on 140 IEEE Explore and on 80 Researchgate. People are busy in their day to day activities so they are making use of the technology to pay the bills, online shopping etc.,. Where there is high scope for credit card fraud activities (Jog and Chandavale 2017) ("Credit Card Fraud Detection System: A Survey" 2020). As usage has increased fradulents are making this as an advantage in credit card thefts and causing big problems for payment gateways and banks (Sherly and Nedunchezhian 2010). Such activities can be stopped by developing machine learning algorithms(Shah et al. 2019).

Our team has extensive knowledge and research experience that has translated into high quality publications (K. Mohan et al. 2022; Vivek et al. 2022; Sathish et al. 2022; Kotteeswaran et al. 2022; Yaashikaa, Keerthana Devi, and Senthil Kumar 2022; Yaashikaa, Senthil Kumar, and Karishma 2022; Saravanan et al. 2022; Jayabal et al. 2022; Krishnan et al. 2022; Jayakodi et al. 2022; H. Mohan et al. 2022). Previous research is using the improper datasets as its input data set("Table S1: Groups of Training and Testing Datasets," n.d.). Positive classes are highly oversampled (Shiny Irene et al. 2020). The number of tuples obtained in the dataset after cleaning of the dataset is very low, so the size of the input dataset is very small (Sherly and Nedunchezhian 2010; Ghosh, Ghosh, and Reilly 1994) . Accuracy of the existing research is very low. Research aims to increase the accuracy of the algorithms and to increase the size of the input dataset(Friedman et al. 2022).

## 2. Materials and Methods

The testing and the study of the dataset has been done in the college laboratory. Implementation, execution, error rectification was done in the university. The project does not require any human samples and human data and the project does not require any ethical and moral permissions. There are totally 2 groups mentioned in the study. One group is mentioned for the Novel Random forest classifier and another group is mentioned for the Linear Regression algorithm. Each group consists of a sample size of 34. The statistical test difference between g-power 0.08 and alpha values are $\alpha$=0.05. The data sets have been taken from wallethub.com.

**Random Forest Classifier**

Novel Random Forest is one of the most important and widely used machine learning algorithms which comes under supervised learning technique. It tends to be utilized for both Classification and Regression issues in Machine Learning. It depends on the idea of ensemble language, which is a course of consolidating various classifiers to take care of a complicated issue and to work on the performance of the algorithm. It chooses the best pattern in means of voting in the last phase (Sherly and Nedunchezhian 2010; Ghosh, Ghosh, and Reilly 1994; Shah et al. 2019)).Rather than depending on one decision tree, the random forest takes the decision from each tree in view of the larger part expectation of the votes and it predicts the result.

**Random Forest pseudocode**

1.Randomly select "X" features from total "Y" features.Where X << Y

2.Among the "X" features, calculate the node "d" using the best split point.

3.Split the node into daughter nodes using the best split.

4.Repeat 1 to 3 steps until "l" number of nodes has been reached.

5.Build forest by repeating steps 1 to 4 for "n" number times to create "n"number of trees.

**Linear Regression**

Linear Regression is a machine learning algorithm in view of managed learning. Regression task is being performed. Regression models an objective expectation esteem in view of independent variables. It is generally utilized for figuring out the connection among variables and estimating.

Regression models are different based on the dependent and independent variables. Based on the kind of relationship between dependent and independent variables they are considering and the number of independent variables being used. Relation between dependent(y) and independent (y) variables is given the Linear Regression (Sherly and Nedunchezhian 2010; Ghosh, Ghosh, and Reilly 1994; Shah et al. 2019; Dal Pozzolo et al. 2018)). Since linear regression shows the linear relationship, change of value of the dependent variable is dependent on the independent variable.

**Linear Regression pseudocode**

1. Start.
2. Read Number of Data (n)
3. For i=1 to n:
Read $X_i$ and $Y_i$
Next i
4. Initialize:
sumX = 0
sumX2 = 0
sumY = 0
sumXY = 0
sumX = sumX + $X_i$
5. Calculate Required Sum
For i=1 to n:
sumX2 = sumX2 + $X_i$ * $X_i$
sumY = sumY + $Y_i$
sumXY = sumXY + $X_i$ * $Y_i$
Next i
6. Calculate Required Constant a and b of y = a + bx:
b = (n * sumXY - sumX * sumY)/(n*sumX2 - sumX * sumX)
a = (sumY - b*sumX)/n
7. Display value of a and b
8. Stop

There are two types of variables: one is a dependent variable and the other is an independent variable. Every variable is unique and has different functions to perform their roles. The independent data are the attributes like id, name, url, status_count and others and the dependent variables are update, transaction_class which consists of binary data. This study uses t test as its testing element.

**Statistical Analysis**

A t-test is a kind of inferential static used to decide whether there is a tremendous contrast between the method for two groups, which might be connected in specific elements. The spss instrument has been utilized to get the result for the grouped statistics. The t-test is one of many tests utilized with the end goal of theory testing in insights. The statistical tool used for this study is the IBM SPSS version. Attributes like update, transaction_class are considered as dependent variables for this study. In SPSS, the datasets are prepared using 34 as sample size for both the classifiers Novel Random Forest and Linear Regression. Groupid is given as 1 for Novel Random Forest and 2 for Linear Regression, the Groupid is given as a grouping variable and accuracy is given as a testing variable.

## 3. Results

In the spss the sample size is given as sample size of 34. For analyzing the process of Novel Random forest and Linear regression above sample data is used. Loss is also calculated for both the algorithms to make the comparison by using these 34 data samples. This data is used for the analysis of the Novel Random forest classifier and Linear Regression. These 34 data samples used for each algorithm along with their loss are also used to calculate statistical values that can be used for comparison (Paasch, n.d.; Deb, Ghosal, and Bose, n.d.). From Table 1, group, accuracy and the loss values for the two algorithms Novel Random Forest classifier and Linear Regression classifier are denoted and classified. Number of samples that are collected and the mean and standard deviation obtained for the accuracies are entered in the group statistics table (Paasch, n.d.)). From Table 2, the group statistics values along with the mean, standard deviation and the standard error mean for the two algorithms are also specified. The Independent sample T test is applied for the data set fixing confidence interval as 95%. Below image specifies the graph that depicts the comparison between Novel Random Forest and Linear Regression based on their accuracy. Table 3 shows the independent t sample test for the algorithms. The comparative accuracy analysis, mean of loss between the two algorithms are specified (Jog and Chandavale 2017)). Fig. 1 shows the comparison of mean of accuracy and mean loss between Novel Random Forest Classifier and Linear Regression algorithm.

## 4. Discussions

From the given study the accuracy of the novel random forest classifier is 94.4% when compared to the accuracy of the Linear Regression is 51.9%. Sample size is given as 30 analyses of statistics have been done for both the Novel Random forest classifier and the Linear Regression algorithms in order to compare both the algorithms to find the better analysis algorithm in fraud detection of credit cards(Benware 2021). For the given classifiers the group and accuracy has been calculated. The mean, standard deviation and the standard mean values for the Novel Random Forest classifier algorithms are 85.4240, 6.12781 and 1.93778 respectively. This clearly indicated that Novel Random Forest is a better classifier when compared to Linear Regression classifier. Linear Regression classifiers, the mean, the standard deviation and the standard mean are 47.9700, 3.48139 and 1.10091 respectively. Contrasted with past investigation of Novel Random forest classifier and Linear Regression algorithms for credit card fraud detection, our examination has got better accuracy in detecting credit card fraud examination. previously the Novel Random forest classifier got the accuracy of 90.2% and for the Linear Regression got the accuracy of 48.5%. But in our analysis we got the accuracy of the Novel Random forest classifier is 94.4% and for the Linear Regression we got the accuracy of 51.9%.Datasets have been ousted from different assets and these datasets may contain a few independent and undesirable credits are there, this ought to be eliminated to get the best accuracy. Hence, the information is tested and prepared to get the best output accuracy (Dal Pozzolo et al. 2018). Testing consumes most of the time(Akinbohun and Atanlogun 2018). Training for this cycle takes a long time. Execution time likewise happens in high time due to the irregular condition of the datasets(Filippov, Mukhanov, and Shchukin 2008) (Shah et al. 2019). The training data can be changed but it is a time consuming process and if data is trained low, the accuracy will be reduced("Credit Card Fraud Detection System: A Survey" 2020). The coordinated based grouping ought to be workable for testing and planning for both the Machine learning Techniques Novel Random Forest and Linear Regression. The data cleaning cycle can be furthermore improved and the time of execution can be lessened("Credit Card Fraud Detection System: A Survey" 2020). The process of everything working out usage in setting up the dataset can be decreased. Fraud recognition can be utilized on credit cards and also E-wallets. Subsequently this novel recognition for cyber threats is particularly useful in recognizing a novel methodology for false rate decrease.

## 5. Conclusion

For this project the data has been collected from various sources for the usage in the credit card fraud detection. The comparison has been made between the Novel Random forest classifier and the Linear Regression for this analysis. The accuracy of the novel random forest classifier is 94.4% and the Linear Regression accuracy of 51.9%.

**Declarations**

**Conflict of Interests**

No conflict of interest in this manuscript.

**Authors Contribution**

Author EMM was involved in data collection, data analysis and manuscript writing. Author JJT was involved in conceptualization, data validation and critical reviews of manuscript.

**Acknowledgements**

The authors would like to express their gratitude towards Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences (Formerly known as Saveetha University) for providing the necessary infrastructure to carry out this work successfully.

**Funding**

We thank the following organizations for providing financial support that enabled us to complete the study.

1.Dreamsplus,Chennai.

2. Saveetha University

3.Saveetha Institute of Medical and Technical Sciences

4.Saveetha School of Engineering

## 6. References

Akinbohun, Folake, and Sunday Kolawole Atanlogun. 2018. "Credit Card Fraud Detection System in Commercial Sites." European Journal of Engineering Research and Science. https://doi.org/10.24018/ejers.2018.3.11.920.

Benware, Dewey. 2021. Machine Learning Algorithms: How The Machine Learning Algorithms Work Behind The Scenes: Random Forest Algorithm Geeksforgeeks.

"Credit Card Fraud Detection System: A Survey." 2020. Journal of Xidian University. https://doi.org/10.37896/jxu14.5/599.

Dal Pozzolo, Andrea, Giacomo Boracchi, Olivier

Caelen, Cesare Alippi, and Gianluca Bontempi. 2018. "Credit Card Fraud Detection: A Realistic Modeling and a Novel Learning Strategy." IEEE Transactions on Neural Networks and Learning Systems 29 (8): 3784–97.

Deb, Koushik, Souvik Ghosal, and Debkanya Bose. n.d. "A Comparative Study on Credit Card Fraud Detection." https://doi.org/10.31224/osf.io/8ctxd.

Filippov, V., L. Mukhanov, and B. Shchukin. 2008. "Credit Card Fraud Detection System." 2008 7th IEEE International Conference on Cybernetic Intelligent Systems. https://doi.org/10.1109/ukricis.2008.4798919.

Friedman, Lee, Vladyslav Prokopenko, Shagen Djanian, Dmytro Katrychuk, and Oleg V. Komogortsev. 2022. "Factors Affecting Inter-Rater Agreement in Human Classification of Eye Movements: A Comparison of Three Datasets." Behavior Research Methods, April. https://doi.org/10.3758/s13428-021-01782-4.

Ghosh, Ghosh, and Reilly. 1994. "Credit Card Fraud Detection with a Neural-Network." Proceedings of the Twenty-Seventh Hawaii International Conference on System Sciences HICSS-94. https://doi.org/10.1109/hicss.1994.323314.

Jayabal, Ravikumar, Sekar Subramani, Damodharan Dillikannan, Yuvarajan Devarajan, Lakshmanan Thangavelu, Mukilarasan Nedunchezhiyan, Gopal Kaliyaperumal, and Melvin Victor De Poures. 2022. "Multi-Objective Optimization of Performance and Emission Characteristics of a CRDI Diesel Engine Fueled with Sapota Methyl Ester/diesel Blends." Energy. https://doi.org/10.1016/j.energy.2022.123709.

Jayakodi, Santhoshkumar, Rajeshkumar Shanmugam, Bader O. Almutairi, Mikhlid H. Almutairi, Shahid Mahboob, M. R. Kavipriya, Ramesh Gandusekar, Marcello Nicoletti, and Marimuthu Govindarajan. 2022. "Azadirachta Indica-Wrapped Copper Oxide Nanoparticles as a Novel Functional Material in Cardiomyocyte Cells: An Ecotoxicity Assessment on the Embryonic Development of Danio Rerio." Environmental Research 212 (Pt A): 113153.

Jog, Anita, and Anjali Chandavale. 2017. "Database Implementation and Testing of Dynamic Credit Card Fraud Detection System." International Journal of Computer Applications. https://doi.org/10.5120/ijca2017914557.

Kotteeswaran, C., Indrajit Patra, Regonda Nagaraju, D. Sungeetha, Bapayya Naidu Kommula, Yousef Methkal Abd Algani, S. Murugavalli, and B. Kiran Bala. 2022. "Autonomous Detection of Malevolent Nodes Using Secure Heterogeneous Cluster Protocol." Computers and Electrical Engineering. https://doi.org/10.1016/j.compeleceng.2022.107902.

Krishnan, Anbarasu, Duraisami Dhamodharan, Thanigaivel Sundaram, Vickram Sundaram, and Hun-Soo Byun. 2022. "Computational Discovery of Novel Human LMTK3 Inhibitors by High Throughput Virtual Screening Using NCI Database." Korean Journal of Chemical Engineering. https://doi.org/10.1007/s11814-022-1120-5.

Mohan, Harshavardhan, Sethumathavan Vadivel, Se-Won Lee, Jeong-Muk Lim, Nanh Lovanh, Yool-Jin Park, Taeho Shin, Kamala-Kannan Seralathan, and Byung-Taek Oh. 2022. "Improved Visible-Light-Driven Photocatalytic Removal of Bisphenol A Using V2O5/WO3 Decorated over Zeolite: Degradation Mechanism and Toxicity." Environmental Research. https://doi.org/10.1016/j.envres.2022.113136.

Mohan, Kannan, Abirami Ramu Ganesan, P. N. Ezhilarasi, Kiran Kumar Kondamareddy, Durairaj Karthick Rajan, Palanivel Sathishkumar, Jayakumar Rajarajeswaran, and Lorenza Conterno. 2022. "Green and Eco-Friendly Approaches for the Extraction of Chitin and Chitosan: A Review." Carbohydrate Polymers 287 (July): 119349.

Paasch, Carsten A. W. n.d. "Credit Card Fraud Detection Using Artificial Neural Networks Tuned by Genetic Algorithms." https://doi.org/10.14711/thesis-b1023238.

Padvekar, Suchita Anant, Atharva College of Engineering Department of Computer Science University of Mumbai, Pragati Madan Kangane, and Komal Vikas Jadhav. 2016. "Credit Card Fraud Detection System." International Journal Of Engineering And Computer Science. https://doi.org/10.18535/ijecs/v5i4.22.

Saravanan, A., P. Senthil Kumar, B. Ramesh, and S. Srinivasan. 2022. "Removal of Toxic Heavy Metals Using Genetically Engineered Microbes: Molecular Tools, Risk Assessment and Management Strategies." Chemosphere 298 (July): 134341.

Sathish, T., R. Saravanan, V. Vijayan, and S. Dinesh Kumar. 2022. "Investigations on Influences of MWCNT Composite Membranes in Oil Refineries Waste Water Treatment with Taguchi Route." Chemosphere 298 (July): 134265.

Shah, Vvyom, Parin Shah, Harshit Shetty, and Kamal Mistry. 2019. "Review of Credit Card Fraud Detection Techniques." 2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN).

https://doi.org/10.1109/icscan.2019.8878853.
Sherly, K. K., and R. Nedunchezhian. 2010. "BOAT Adaptive Credit Card Fraud Detection System." 2010 IEEE International Conference on Computational Intelligence and Computing Research. https://doi.org/10.1109/iccic.2010.5705824.
Shiny Irene, D., V. Surya, D. Kavitha, R. Shankar, and S. John Justin Thangaraj. 2020. "An Intellectual Methodology for Secure Health Record Mining and Risk Forecasting Using Clustering and Graph-Based Classification." Journal of Circuits, Systems and Computers, November. https://doi.org/10.1142/S0218126621501358.
"Table S1: Groups of Training and Testing Datasets." n.d. https://doi.org/10.7717/peerj.4380/supp-2.
Vivek, J., T. Maridurai, K. Anton Savio Lewise, R. Pandiyarajan, and K. Chandrasekaran. 2022. "Recast Layer Thickness and Residual Stress Analysis for EDD AA8011/h-BN/B4C Composites Using Cryogenically Treated SiC and CFRP Powder-Added Kerosene." Arabian Journal for Science and Engineering. https://doi.org/10.1007/s13369-022-06636-5.
Yaashikaa, P. R., M. Keerthana Devi, and P. Senthil Kumar. 2022. "Algal Biofuels: Technological Perspective on Cultivation, Fuel Extraction and Engineering Genetic Pathway for Enhancing Productivity." Fuel. https://doi.org/10.1016/j.fuel.2022.123814.
Yaashikaa, P. R., P. Senthil Kumar, and S. Karishma. 2022. "Review on Biopolymers and Composites – Evolving Material as Adsorbents in Removal of Environmental Pollutants." Environmental Research. https://doi.org/10.1016/j.envres.2022.113114.

**Tables and Figures**

Table 1. Group, Accuracy and Loss value calculation for Random Forest and Linear Regression classifiers.

| | Name | Type | Width | Decimals | Columns | Measure | Role |
|---|---|---|---|---|---|---|---|
| 1 | Group | Numeric | 8 | 0 | 70 | Nominal | Input |
| 2 | Accuracy | Numeric | 8 | 4 | 70 | Scale | Input |
| 3 | Loss | Numeric | 8 | 2 | 70 | Scale | Input |

Table 2: Statistical analysis of mean, standard deviation and standard error of precision and accuracy of Random Forest Classifier and Linear Regression Classifier.

| | Group | Algorithm | N | Mean | Std.Deviation | Std.Error Mean |
|---|---|---|---|---|---|---|
| Accuracy | 1 | Random Forest | 68 | 85.4240 | 6.12781 | 1.93778 |
| | 2 | Linear Regression | 68 | 47.9700 | 3.48139 | 1.10091 |
| Loss | 1 | Random Forest | 68 | 14.5760 | 6.12781 | 1.93778 |
| | 2 | Linear Regression | 68 | 52.0300 | 3.48139 | 1.10091 |

Table 3: Comparison of significance level for Random Forest Classifier and Linear Regression Classifier with value $p < 0.05$.

| | | F | sig. | t | df | sig.(2-tailed) | Mean Difference | Std,Error difference | Lower | Upper |
|---|---|---|---|---|---|---|---|---|---|---|
| Accuracy | Equal | 5.310 | .033 | 16.805 | 18 | .00 | 37.45400 | 2.22868 | 32.77172 | 42.13628 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Variance assumed | | | | | | | | | |
| | Equal Variance not assumed | | | 16.805 | 14.262 | .00 | 37.45400 | 2.22868 | 32.68217 | 42.22583 |
| Loss | Equal Variance assumed | 5.310 | .033 | -16.805 | 18 | .00 | -37.45400 | 2.22868 | -42.13628 | -32.77172 |
| | Equal Variance not assumed | | | -16.805 | 14.262 | .00 | -37.45400 | 2.22868 | -42.22583 | -32.68217 |

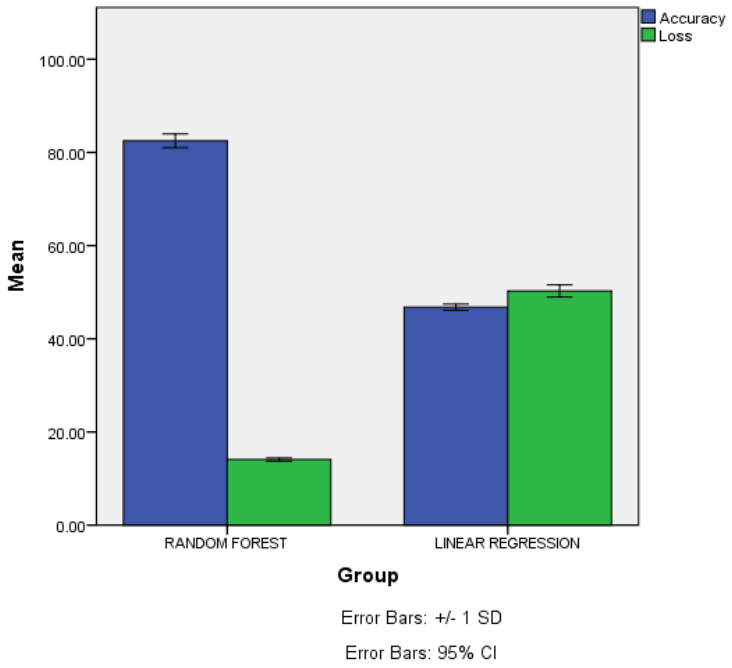


Fig. 1: Comparison of Random Forest and Linear Regression Classifier in terms of mean accuracy. The mean accuracy of Random Forest is better than Linear Regression. The standard deviation of Random Forest is slightly better than Linear Regression X Axis: Random Forest vs Linear Regression Classifier Y Axis: Mean accuracy of detection ± 1SD.