Original Article

Comparison of Novel Optimized Random Forest Technique and Gradient Boosting for Credit Card Fraud Detection with Improved Precision

M.Shahid Saif Ali Baig[1], K.Jaisharma[2]

[1]Research Scholar, Department of Computer Science And Engineering, Saveetha School of Engineering, Saveetha Institute of Medical And Technical Sciences, Saveetha University, Chennai, Tamil Nadu, India. Pincode: 602 105.
[2]Project Guide, Department of Computer Science And Engineering, Saveetha School of Engineering, Saveetha Institute of Medical And Technical Sciences, Saveetha University, Chennai, Tamil Nadu, India. Pincode: 602 105.

Abstract

Aim: The purpose of this research study is to detect credit card fraud using Novel Optimized Random Forest Technique (NORFT) and Gradient Boosting (GB). **Materials and Methods:** The Novel Optimized Random Forest Technique Algorithm uses parallel Decision Tree technique in addition with Random Forest Technique to improve the prediction of Credit Card Fadulants. Total sample size of 40 is used for testing and analysis, based on Gpower statistical analysis tool by considering gpower 0.8. In NORFT used N=20 and in GB used N=20 to measuring the performance of both algorithms. **Result:** Novel Optimized Random Forest Technique provides mean precision of 92.52%, and compared with Gradient Boosting algorithm of mean precision is 88.56%. Statistical significance value was fixed as ($p>0.05$) and obtained 0.477, this shows that NORFT is not statistically significant with alternative hypotheses. **Conclusion:** Based on the result, improved precision comparison results show that the efficiency of the Novel Optimized Random Forest Technique is better than Gradient Boosting.

Keywords: Anomaly Detection, Credit Card Fraudulent, Internet Security, Gradient Boosting, Machine Learning, Novel Optimized Random Forest Technique.

DOI: 10.47750/pnr.2022.13.S04.0100

INTRODUCTION

Crime rate in digital form keeps on increasing day by day, by targeting individuals, companies and governments organizations. Using digital crime, illegal activities such as stealing users credentials, crossing law boundaries, child poronography, tracking user information, etc. One such fraudulent digital crime is stealing credit card information and using it for illegal purposes without the knowledge of the users (Paasch, n.d.). The importance of this study is to address credit card fraud. Anomaly detection is to find any dubious activities in the transaction which are occurring using credit cards. This process of detection involves creation of reports with all the data combined that has taken place in the transaction process and transferred to a certain authority whose transactions seem doubtful. The other transaction will be processed automatically if it is found decisive. Finding the anomalies in the transaction data and descriptive analysis with internet security. Cyber crimes related to online money transfers are increasing nowadays. The main fraud revolves around a transaction that uses cards. Credit card frauds have been increasing day by day. Frauds are taking place using different means like internet phone calls etc (Kiruthika et al. 2020). The applications of credit card fraud detection techniques help many people to save their money from the attackers and various online hackers (Garg, Chaudhary, and Mishra 2021). It can be used for preventing hackers and providing internet security to users. It can be used as an online tool to prevent fraud from attackers (Lamba 2020).

In this study, related to credit card fraud, published articles were taken in the last five years from the two databases namely (IEEE-29, ScienceDirect-110). Articles which are mostly cited are Visa card extortion with the help of unsupervised data to secure the user privacy information (Rai and Dwivedi 2020). In this paper Execution about Examination to predictive analytics for proceeding in visa cards extortion of fraud identification were analyzed (Jain, Agrawal, and Kumar 2020). This paper contains, in its original methodology, a master card approach to credit cards with two algorithms (Azhan and Meraj 2020). The authors (Dileep, Navaneeth, and Abhishek 2021), discussed fake detection methods in credit card transactions (Shmatko, Fedorchenko, and Prochukhan 2021). Among all the articles, the best paper identified using high citations are Fraud Anomaly

Detection Techniques in Visa information utilizing closely supervised ML based schemes and this taken as base for this research work (Goyal and Sharma 2020). Our team has extensive knowledge and research experience that has translate into high quality publications(Bhansali et al. 2021; Jayanth et al. 2021; Sudhakar, Ravel, and Perumal 2021; Sathiyamoorthi et al. 2021; Deepanraj et al. 2021; Raju et al. 2021; Arun Prakash et al. 2020; Kamath et al. 2020; Shanmugam et al. 2021; Rajasekaran et al. 2020; Adhinarayanan et al. 2020; Rajesh et al. 2020; Aurtherson et al. 2021)

In the existing model, the precision is not more than 88.5% and this may lead to wrong classification during digital fraud detection which causes theft of personal values. The proposed model aim is to improve precision more than the existing research and experiment the research in minimal computational cost using a machine learning algorithm named Novel Optimized Random Forest Technique (NORFT).

Materials And Methods

The setup of this research had been performed in the Data Analytics Laboratory of Department Computer Science and Engineering in Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences. The Research carries with the two algorithms one is for base and another for comparison. The algorithms taken as two groups, which was group 1 as Novel Optimized Random Forest Technique (NORFT) and group 2 was Gradient Boosting (GB) with two set of sample size of 20, which is total of 40 which is performed under pretest power of 80% (Dileep, Navaneeth, and Abhishek 2021).

Novel Optimized Random Forest Technique

Random Forest (RF) is an extremely valuable and useful ML algorithm. During training the Random Forest model makes numerous choices of tree structures. RF is a regulated learning approach which needs test information for the model for preparing. It makes arbitrary random forests for the issue set and afterwards finds the arrangement utilizing these arbitrary random forests. One of the most essential functions of the RF technique is that it may handle the statistics set containing non stop variables. In the case of registration and express variables as inside the case of class classifications. The Mean Squared Error (MSE) was calculated using Equation 1 in the NORFT. It performed with better results for categorization problems, by combining the power of Dual Terminal and RF(Dileep, Navaneeth, and Abhishek 2021).

$$MSE=1N\ i=1N\ (fi\text{-}yi)^2 \quad (1)$$

where, N is defined as the number of significant pieces of information data of a point, MSE is defined as Mean Squared Error, fi is the worth returned by the model, yi is the genuine incentive for information point I.

Gradient Boosting

Gradient Boosting (GB) algorithm brings out a succession of decision trees by exercising them. GB is used to set classification and regression at identical times. At each phase a newly decision tree is constructed on the grounds of preceding decision tree which helps in minimisation of flaws. GB takes an additional period of time and memory for suitable algorithms to evaluate and procession a dataset. Its count on the induction when the possible finest succeeding model is integrated with preceding models reduces the overall prediction error. The mathematical statement for GB appeared in the Equation 2 (Kiruthika et al. 2020).

$$f(x) =\arg\min \gamma\ L(y,\gamma) \quad (2)$$

The model was built on setup with the hardware components used were i3 processor, 4GB RAM and 1TB hard disk by using HP system. The softwares used was Windows 11, Google Colab. The Charge card Fraud dataset downloaded from Kaggle website, preprocessed the dataset as required to train the model gives the results. The dataset used has 284807 credit card related transaction information with unique attributes along the timestamp from v1 to v8.

Statistical Analysis

IBM SPSS version 22 software is used for statistical analysis of Novel Optimized Random Forest Technique (NORFT) and Gradient Boosting (GB) based methods. The independent variables are Time, User identities, Amount, Sensitive features (V1 TO V28) and Dependent variables are prediction of precision 1-fraud,0-Otherwise. The dependent variable was precision. Statistical analysis was carried out using an independent sample T-test (Dileep, Navaneeth, and Abhishek 2021).

Results

Table 1 shows the outcome of group statistics models by comparing the algorithms precision for the sample values=20 for Novel Optimized Random Forest Technique (NORFT) and mean value is 92.5200 and standard deviation=.72296 and standard error mean is .16166. Similarly for Gradient Boosting (GB) Algorithm N=20 samples taken, mean value is=88.5660 and standard deviation=.81780 and standard error mean is 0.18287. It was observed that the mean precision of the Novel Optimized Random Forest Technique algorithm (NORFT) was 92.5200 and the Gradient Boosting algorithm was 88.5660.

Table 2 shows the outcome of the Independent sample T-test performed to compare the precision of Novel Optimized Random Forest Technique (NORFT) with Gradient Boosting (GB) algorithms. The significance 0.477 ($p<0.05$) with confidence interval is 95%.

The graphical comparison of both algorithms shown in Figure. 1, the bar was drawn by keeping NORFT and GB along with error bars on top in X-axis and Mean precision in Y-axis by considering +/-1SD error bar with 95% Confidence Interval (CI). It is observed that a Novel Optimized Random Forest Technique (92.5200) has a higher precision with less error bars, when compared to the Gradient Boosting algorithm (88.5660).

Discussion

The Novel Optimized Random Forest Techniques (NORFT) and Gradient Boosting algorithms (GB) are implemented and compared their performance using credit card fraud dataset for the prediction. From obtained results, significance attained is 0.477 ($p>0.05$) has no significance due to dataset inconsistency. Novel Optimized Random Forest Technique (NORFT) provides precision 92.52% and Gradient Boosting algorithm (GB) is 88.56%. Based on the precision percentage, it is evidence that NORFT is better than GB.

There are many articles similar to the study of proposed models where findings are an original methodology for Mastercard extortion location utilizing choice tree and arbitrary backwoods calculations by internet security (Chowdari and Bhargav Chowdari 2021).In this aspect prefer to Naive Bayes (NB), and combination with two algorithm for credit fraud detections (Saheed et al. 2020). Peculiarity of perception fraud observation was using find a Random forest (RF) in mastercards (Xuan et al. 2018). Visa fraud detection system uses network security with the help of Random Forest and GBiers of network layer (Sudha and Akila 2021). Opposite algorithm for the proposed study is mastercard exchanges of data Transactions is more accurate in deep learning Domain (Shao, Gu, and Zhang 2020), which has a precision of 18.30% compared to Novel Optimized Random Forest Technique (NORFT) which has 83.50% of precision.

This proposed model has limitations to poor performance with incomplete dataset, because credit card records contain more sensitive data. So, it's not possible to provide all details of users confidential data, this results in poor performance for few uncertain situations in Credit Card Fraudulent. This study, in the future, can be used for various scenarios like stock market prediction, disease diagnoses, and predictions using precision, it will be helpful for numerous other applications in the future.

Conclusion

The precision of Novel Optimized Random Forest Technique (NORFT) has been improved to (92.52%) compared to Gradient Boosting algorithm (GB) which is having (88.56%). This, also concluded with statistical significance and other supporting related articles, proves the Novel Optimized Random Forest Technique (NORFT) is an efficient algorithm when compared to Gradient Boosting algorithm (GB).

DECLARATIONS

Conflict of Interest

No conflict of interests in this manuscript.

Authors Contribution

Author MSSAB was involved in dataset collection, algorithm development and manuscript writing. Author KJS was involved in validation and review of the manuscript.

Acknowledgement

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

Fundings

We would like to thank the following organization for providing financial support that enabled us to complete this study.

1.Biozone Pvt.Ltd, Chennai.
2. Saveetha university.
3. Saveetha Institute of Medical and Technical Sciences.
4. Saveetha School of Engineering.

References

1. Adhinarayanan, Rajesh, Aravindh Ramakrishnan, Gopal Kaliyaperumal, Melvinvíctor De Poures, Rajesh Kumar Babu, and Damodharan Dillikannan. 2020. "Comparative Analysis on the Effect of 1-Decanol and Di-N-Butyl Ether as Additive with diesel/LDPE Blends in Compression Ignition Engine." *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects*, June, 1–18.
2. Arun Prakash, V. R., J. Francis Xavier, G. Ramesh, T. Maridurai, K. Siva Kumar, and R. Blessing Sam Raj. 2020. "Mechanical, Thermal and Fatigue Behaviour of Surface-Treated Novel Caryota Urens Fibre–reinforced Epoxy Composite." *Biomass Conversion and Biorefinery*, August. https://doi.org/10.1007/s13399-020-00938-0.
3. Aurtherson, P. Babu, Bhanu Teja Nalla, Karthikeyan Srinivasan, Kulmani Mehar, and Yuvarajan Devarajan. 2021. "Biofuel Production from Novel Prunus Domestica Kernel Oil: Process Optimization Technique." *Biomass Conversion and Biorefinery*, May. https://doi.org/10.1007/s13399-021-01551-5.
4. Azhan, Mohammed, and Shazli Meraj. 2020. "Credit Card Fraud Detection Using Machine Learning and Deep Learning Techniques." *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)*. https://doi.org/10.1109/iciss49785.2020.9316002.
5. Bhansali, Karan J., Kamlesh R. Balinge, Subodh U. Raut, Shubham A. Deshmukh, M. Senthil Kumar, C. Ramesh Kumar, and Pundlik R. Bhagat. 2021. "Visible Light Assisted Sulfonic Acid-Functionalized Porphyrin Comprising Benzimidazolium Moiety for Photocatalytic Transesterification of Castor Oil." *Fuel* 304 (November): 121490.
6. Chowdari, G. Bhargav, and G. Bhargav Chowdari. 2021. "SUPERVISED MACHINE LEARNING ALGORITHMS FOR DETECTING CREDIT CARD FRAUD." *EPRA International Journal of Research & Development (IJRD)*. https://doi.org/10.36713/epra7636.
7. Deepanraj, B., N. Senthilkumar, D. Mala, and A. Sathiamourthy. 2021. "Cashew Nut Shell Liquid as Alternate Fuel for CI Engine—optimization Approach for Performance Improvement." *Biomass Conversion and Biorefinery*, February. https://doi.org/10.1007/s13399-021-01312-4.
8. Dileep, M. R., A. V. Navaneeth, and M. Abhishek. 2021. "A Novel Approach for Credit Card Fraud Detection Using Decision Tree and Random Forest Algorithms." *2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)*. https://doi.org/10.1109/icicv50876.2021.9388431.
9. Garg, Vaishali, Sarika Chaudhary, and Anil Mishra. 2021. "ANALYSING AUTO ML MODEL FOR CREDIT CARD FRAUD DETECTION." *International Journal of Innovative Research in Computer Science & Technology*. https://doi.org/10.21276/ijircst.2021.9.3.5.
10. Goyal, Yogita, and Anand Sharma. 2020. *Credit Card Fraud Detection and Analysis Through Machine Learning*.
11. Jain, Vinod, Mayank Agrawal, and Anuj Kumar. 2020. "Performance Analysis of Machine Learning Algorithms in Credit Cards Fraud Detection." *2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*. https://doi.org/10.1109/icrito48877.2020.9197762.
12. Jayanth, Bellappu Venkat, Melvin Victor Depoures, Gopal Kaliyaperumal, Damodharan Dillikannan, Dilipsingh Jawahar, Kumaran Palani, and Ganesha Prasad Meravanigee Shivappa. 2021. "A Comprehensive Study on the Effects of Multiple Injection Strategies and Exhaust Gas Recirculation on Diesel Engine Characteristics That Utilize Waste High Density Polyethylene Oil." *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects*, June, 1–18.
13. Kamath, Manjunath, Subha Krishna Rao, Jaison, Sridhar, Kasthuri, Gopinath, Sivaperumal, and Shantanu Patil. 2020. "Melatonin Delivery from PCL Scaffold Enhances Glycosaminoglycans Deposition in Human Chondrocytes – Bioactive Scaffold Model for Cartilage Regeneration." *Process Biochemistry* 99 (December): 36–47.
14. Kiruthika, Usha, S. Kanaga Suba Raja, C. J. Raman, and V. Balaji. 2020. "A Novel Fraud Detection Scheme for Credit Card Usage Employing Random Forest Algorithm Combined with Feedback Mechanism." *2020 International Conference on Power, Energy, Control and Transmission Systems (ICPECTS)*. https://doi.org/10.1109/icpects49113.2020.9337045.
15. Lamba, Harshit. 2020. *Credit Card Fraud Detection In Real-Time*.
16. 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.
17. Rai, Arun Kumar, and Rajendra Kumar Dwivedi. 2020. "Fraud Detection in Credit Card Data Using Machine Learning Techniques." *Communications in Computer and Information Science*. https://doi.org/10.1007/978-981-15-6318-8_31.
18. Rajasekaran, S., D. Damodharan, K. Gopal, B. Rajesh Kumar, and Melvin Victor De Poures. 2020. "Collective Influence of 1-Decanol Addition, Injection Pressure and EGR on Diesel Engine Characteristics Fueled with diesel/LDPE Oil Blends." *Fuel* 277 (October): 118166.
19. Rajesh, A., K. Gopal, De Poures Melvin Victor, B. Rajesh Kumar, A. P. Sathiyagnanam, and D. Damodharan. 2020. "Effect of Anisole Addition to Waste Cooking Oil Methyl Ester on Combustion, Emission and Performance Characteristics of a DI Diesel Engine without Any Modifications." *Fuel* 278 (October): 118315.
20. Raju, P., K. Raja, K. Lingadurai, T. Maridurai, and S. C. Prasanna. 2021. "Glass/Caryota Urens Hybridized Fibre-Reinforced nanoclay/SiC Toughened Epoxy Hybrid Composite: Mechanical, Drop Load Impact, Hydrophobicity and Fatigue Behaviour." *Biomass Conversion and Biorefinery*, March. https://doi.org/10.1007/s13399-021-01427-8.
21. Saheed, Yakub K., Moshood A. Hambali, Micheal O. Arowolo, and Yinusa A. Olasupo. 2020. "Application of GA Feature Selection on Naive Bayes, Random Forest and SVM for Credit Card Fraud Detection." *2020 International Conference on Decision Aid Sciences and Application (DASA)*. https://doi.org/10.1109/dasa51403.2020.9317228.
22. Sathiyamoorthi, Ramalingam, Gomathinayakam Sankaranarayanan, Dinesh Babu Munuswamy, and Yuvarajan Devarajan. 2021. "Experimental Study of Spray Analysis for Palmarosa Biodiesel-diesel Blends in a Constant Volume Chamber." *Environmental Progress & Sustainable Energy* 40 (6). https://doi.org/10.1002/ep.13696.
23. Shanmugam, Rajasekaran, Damodharan Dillikannan, Gopal Kaliyaperumal, Melvin Victor De Poures, and Rajesh Kumar Babu. 2021. "A Comprehensive Study on the Effects of 1-Decanol, Compression Ratio and Exhaust Gas Recirculation on Diesel Engine Characteristics Powered with Low Density Polyethylene Oil." *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects* 43 (23): 3064–81.
24. Shao, Mingtian, Naijie Gu, and Xiaoci Zhang. 2020. "Credit Card Transactions Data Adversarial Augmentation in the Frequency Domain." *2020 5th IEEE International Conference on Big Data Analytics (ICBDA)*. https://doi.org/10.1109/icbda49040.2020.9101344.
25. Shmatko, Olexander, Volodimir Fedorchenko, and Dmytro Prochukhan. 2021. "DETECTING CREDIT CARD FRAUD USING MACHINE LEARNING ALGORITHMS." *InterConf*. https://doi.org/10.51582/interconf.19-20.08.2021.037.
26. Sudha, C., and D. Akila. 2021. "Credit Card Fraud Detection System Based on Operational & Transaction Features Using SVM and Random Forest Classifiers." *2021 2nd International Conference on Computation, Automation and Knowledge Management (ICCAKM)*. https://doi.org/10.1109/iccakm50778.2021.9357709.

27. Sudhakar, M. P., Merlyn Ravel, and K. Perumal. 2021. "Pretreatment and Process Optimization of Bioethanol Production from Spent Biomass of Ganoderma Lucidum Using Saccharomyces Cerevisiae." *Fuel* 306 (December): 121680.
28. Xuan, Shiyang, Guanjun Liu, Zhenchuan Li, Lutao Zheng, Shuo Wang, and Changjun Jiang. 2018. "Random Forest for Credit Card Fraud Detection." *2018 IEEE 15th International Conference on Networking, Sensing and Control (ICNSC)*. https://doi.org/10.1109/icnsc.2018.8361343.

TABLES AND FIGURES

Table 1. Novel Optimized Random Forest Technique (92.5200) method and grouped statistics were compared using group statistics for recorded data from simulation for 20 iterations per group Gradient Boosting (GB) (88.5660). In comparison, the Novel Optimized Random Forest Technique (NORFT) algorithm has a high level of precision.

	Algorithm	**N**	**Mean**	**Std. Deviation**	**Std. Error Mean**
Precision	NORFT	20	92.5200	.72296	.16166
	GB	20	88.5660	.81780	.18287

Table 2. Independent sample T test is applied for data set confidence interval is 95% and the level of significance is p=0.316 (p>0.05).

		Levene's Test for Equality of Variance		**T-test for Equality of Means**						
		f	**Sig**	**t**	**df**	**Sig.(2-tailed)**	**Mean Difference**	**Std.Error Difference**	**95% Confidence of the Differences**	
									Lower	**Upper**
Precision	Equal variances assumed	.316	.477	16.200	38	.000	3.95400	.24408	3.45989	4.44811
	Equal variances not assumed			16.200	37.437	.000	3.95400	.24408	3.4565	4.44835

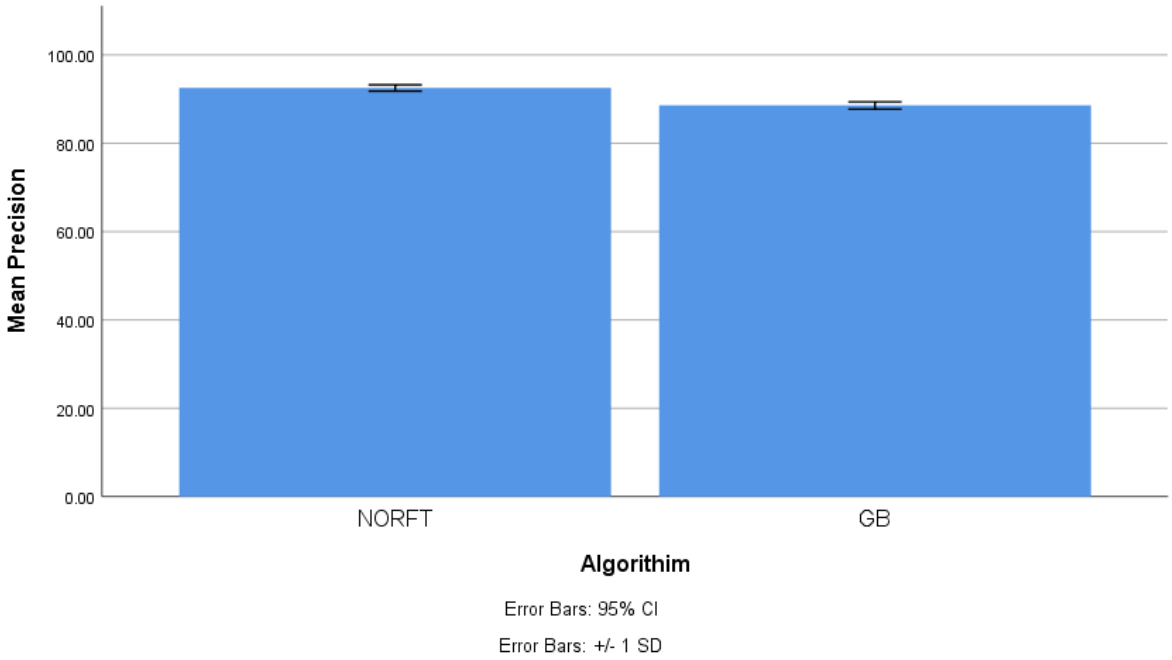

Fig. 1. Performance comparison between Novel Optimized Random Forest Technique is (92.5200%) and Gradient Boosting Algorithm (88.5660%). The mean precision of Novel Optimized Random Forest Technique (NORFT) is better than Gradient Boosting Algorithm (GB). X-axis contains Novel Optimized Random Forest Technique (NORFT) Vs Gradient Boosting Algorithm (GB) algorithm, Y axis Mean precision, where Error Bars +/-1 SD and Error Bars 95% CI.