A Comparison of Logistic Regression Classifier and Random Forest Classifier for the Accurate Classification of Credit Card Fraudulent Transactions

K. Lavanya[1], Rama Parvathy L[2*]

[1]Research Scholar, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University,Chennai,Tamilnadu, India,Pincode:602105

[2*]Project Guide,Corresponding Author, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University,Chennai,Tamilnadu, India,Pincode:602105

ABSTRACT

Aim: The Aim of the research work is credit card fraud detection using random forest classifier in comparison with logistic regression.**Materials and Methods:**novel random forest classifier algorithm with sample size =09 and novel logistic regression algorithm with sample size =09 were evaluated many times to predict the efficiency percentage.The G power taken as 0.8 and a =0.05.**Results and Discussion:**Random forest classifer has been efficiency(97.12%) when compared to logistic regression algorithm efficiency(95.34%). The statistical significance difference(two-tailed) is 0.001($p<0.05$).Based on the above it is observed that the Random forest classifier with 97.12% has better accuracy than 95.34% Logistic regression in Credit card fraud detection.There is a statistical 2-tailed significant difference in accuracy for algorithms is 0.02($p<0.05$) by independent t-test.**Conclusion:**Random forest classifier algorithm significantly better than logistic regression algorithm.

Keywords: Machine Learning, Novel Fraud Transactions, Random Forest Classifier, Logistic Regression, Local outlier factor, Automated Fraud detection.

INTRODUCTION

Fraudsters are getting cash or merchandise or service from unethical and illegal ways which are growing all around the world, today.Considering the expansion of contemporary industry,the likelihood of fraud and money abuse are even a lot.Fraud detection involves observance of the activities of populations of users in order to estimate, understand or avoid objectionable behavior, that include fraud, intrusion, and defaulting.(Dileep, Navaneeth, and Abhishek 2021)Due to the high range of transactions, there's no risk of on-line watching by observers and use of an intelligent system for watching these transactions.As mentioned before, fraudsters attempt to abuse user's banking info and account victimisation, counterfeit credit cards,hacking email, sign and user's account,stealing mastercard number and alternative ways.By improving the accuracy of the classification models,Random forest classifier achieved an overall accuracy of 96.67% and Logistic regression is not that outstanding(Dal Pozzolo et al. 2018).In our projected system we engineered the master card fraud detection victimization novel Machine learning.Such a high

action frequency,credit cards became the primary targets of fraud(Kaur and Kumar 2020).This drawback is especially difficult from the angle of learning, because it is characterized by varied factors such as category imbalance. The range of valid transactions come from deceitful ones. Also, the dealings patterns typically modify their applied math properties over the course of your time(Bahnsen et al. 2013).The key objective of any mastercard fraud detection system is to spot suspicious events and report them to an analyst whereas holding traditional transactions be mechanically processed. For years, money establishments have entrusted this task to rule-based systems that use rule sets written by consultants.

More than 270000 cases are filed during 2019 and 2020 related to the credit card fraud detection.(Kulkarni 2019)This issue conjointly ends up in poor detection accuracy.Third the conception of drift means the behavior of the shopper changes,leading to changes within the information stream once dealing with on-line information detection in real time(Kulkarni 2019).This proposed method has an accuracy of 99% by using novel Random forest algorithm with mentioned approach.(Dileep, Navaneeth, and Abhishek 2021)The algorithm achieved a maximum accuracy of 95% with the mentioned above approach.Using novel machine learning strategies approaches to attenuate the warning rates and increase the fraud detection rate.Using the Proposed system,the accuracy of the algorithm is 99% and 95%.(Zhang, Bhandari, and Black 2020).
(Bhavikatti et al. 2021; Karobari et al. 2021; Shanmugam et al. 2021; Sawant et al. 2021; Muthukrishnan 2021; Preethi et al. 2021; Karthigadevi et al. 2021; Bhanu Teja et al. 2021; Veerasimman et al. 2021; Baskar et al. 2021)

There was a gap in the already existing systems which deal with the novel random forest classifier and novel logistic regression algorithm.Since it may be difficult to predict the fraud transactions and detecting the various novel fraudulent transactions(Kulkarni 2019).It is important to add more values to the dataset and trained dataset predict accurately.(Xuan et al. 2018)As an author, the Brause et al showed however advanced data processing techniques and a novel random forest classifier rule will be combined successfully tested on real mastercard date.Xuan et al (V and Gokula 2019)used 2 forms of random forests ways to coach the behavior options of traditional and abnormal transactions.Authors((T. Kumar 2021) lay the negative choice algorithm on the cloud platform victimization apache Hadoop and MapReduce decreasing the coaching time of basic algorithms.(Kiruthika et al. 2020)This research is to detect the novel fraud transactions with better accuracy using the novel Random forest classifier and Logistic regression(M. S. Kumar et al. 2019).

MATERIALS AND METHODS

This study was conducted in the novel machine learning Laboratory, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences. In this research, two groups were taken in our study:one group refers to the novel random forest classifier and the other group refers to novel logistic regression.The proposed system has 20 samples from the dataset to calculate the software effort estimation

using a clinical calculator. The algorithm random forest classifier and novel logistic regression used in this project. With the G-power having 96% alpha and beta values with 0.05,0.2 respectively and the threshold value is 0.02 and significance value is 0.03.

The dataset collection is taken from an open source kaggle.com for credit card fraud detection using Random forest classifier and Logistic regression.The dataset contains two rows named Time and a variable containing 31 columns and 270000 rows. The dataset is divided into test and train data. The Jupyter software with windows 10.1 system has been to develop this credit card fraud detection.

Random Forest Classifier

Random forest may be a variety of supervised novel machine learning algorithm supported ensemble learning.Ensemble learning is a type of learning wherever you be part of different kinds of algorithms multiple rule of identical sort i.e. multiple decision trees,leading to a forest of trees,therefore the name "Random Forest".The random forest rule are often used for each regression and classification tasks.

PSEUDOCODE

INPUT:Training dataset()

OUTPUT:Accuracy

Step1:Import the packages and import the data set into the program using pandas.

Step2:Convert the string values in the dataset to numerical values.

Step3:Unique values are taken and not any number values are neglected.

Step4:Assign the data to X_train,X_test,y_train,y_test.

Step5:Using train_test_split() function, pass the training and testing variable and give test_size and size as the parameters.

Step6: Important the Random forest classifier from sklearn library.Using the Random forest Classifier,predict the output of the testing variables.

Step7:Display a few digits.Plot the heap map of every pixel in the dataset.

Step8:Calculate the accuracy and plot the confusion matrix for visualizing the number of times the algorithm recognizes the digits.

Logistic Regression

Logistic regression is simpler to implement than Random forest classifier and is extremely economical to coach. It makes no assumptions regarding the distributions of classes within the feature area.It will simply be extended to multiple categories(multinomial regression).it's terribly economical for classifying unknown records.Since the cross-validation technique divides the information into 10 parts, there area unit ten testing knowledge sets. Every testing knowledge set is used to take a look at one classifier (there are ten classifiers). This successively gives the model a bonus by permitting it to use the entire database for testing also as for coaching. The testing method is tightly in addition to the accuracy of the model. Shrewd the final accuracy involves shrewd the accuracy of every classifier.

PSEUDOCODE

INPUT: Training dataset()

OUTPUT: Accuracy

Step1: Import the packages and import the dataset using pandas into the program.

Step2: Split dataset into test(x_test,y_test) and train x_train,y_train).

Step3: Transform the dataset into(n,depth).

Step4: Convert data type of the into float 32 and normalize the data values to another range.

Step5: Increase the sample size and random state,number of filtrations to increase the accuracy and reduce the loss.

Step6: Plot the graph on training accuracy and training loss.Plot the confusion matrix to visualize the predicted values and actual values.

Step7: Plot the numbers and calculate the accuracy and precision.

In the proposed system the training and testing of the data is made in the Jupyter notebook and having used the SPSS software to predict the graph and also G-power software to calculate and pretest for the algorithm to get better percentage of the algorithm.In this proposed system 50 gb hard disk and 8 gb RAM is used for execution of the algorithm.The system type used was a 64-bit OS, X64 and the operating system in windows.

Statistical Analysis

The analysis done by IBM SPSS version 23 for both proposed and existing algorithm iteration were done with the 20 samples and for each iteration the predicted accuracy was noted for analyzing accuracy with value obtained from the iterations. Independent Sample T-test was performed.(Kaur and Kumar 2020)Independent variables are Time which are there in the dataset for prediction and dependent variables are input text for prediction.

RESULTS

In Table 1, it was observed that the Random forest classifier is significantly better than Logistic regression.In the Random forest classifier, the dataset observed that the accuracy and performance of Random forest classifier is better than Logistic regression.In descriptive statistics the accuracy and algorithm values contain the upto 20 values.Standard deviation of Accuracy of Random forest classifier is 97.12% and accuracy of Logistic regression is 95.34%.

In Table 2, the group statistics of algorithms of both Random forest classifier and Logistic regression.Number of Random forest classifiers are 10 and Logistic regressions are 10. Mean of Random forest classifier value is 98.1470 and Logistic regression is 96.9190 and Standard deviation of both the algorithms are 6.1217 and 5.7113.Standard error 1.9359 and 1.8061.In table 2 the two tailed significance values smaller than 0.002($p<0.05$) showed that our hypothesis holds good.

In Figure 1, the independent sample test accuracy equal variance of sig value is 0.034 and the equal variance not assumed of sig value is null.From Figure 1, both the algorithm of Random forest classifier and Logistic regression technique the accuracy value of the Random forest classifier model accuracy is 97.12% and Logistic regression is 95.34%.

DISCUSSION

Based on the above it is observed that the Random forest classifier with 97.12% has better accuracy than 95.34% Logistic regression in Credit card fraud

detection.There is a statistical 2-tailed significant difference in accuracy for algorithms is 0.02($p<0.05$) by independent t-test.

In the existing system the accuracy for the Random forest classifier and Logistic regression are 97% and 95.3% respectively(Sudha and Akila 2021).The accuracy of Random forest classifier and Logistic regression is 98.034% and 98.54% respectively (Sudha and Akila 2021; Alenzi and Nojood 2020).This analysis makes use of novel machine learning to predict the accuracy of credit card fraud detection.The accuracy values for Random forest classifier are 93% and 96%.(T. Kumar 2021) and compared with another model of credit card fraud detection for Random forest classifier and Logistic regression is 99.34% and 94.23%.(Hussein et al. 2021). The Factors affecting the algorithm are sample size of the dataset and test size of the dataset Based on the above finding the existing algorithm was chosen to improve the accuracy.

The limitations of logistic regression is quantity of observations is lesser than the quantity of options,supplying regression shouldn't be used, otherwise,it's going to result in overfitting .It constructs linear boundaries.The major limitation of supplying regression is that the assumption of dimensionality between the variable quantity and therefore the freelance variables.It will solely be accustomed predict separate functions.Hence, the variable quantity of supplying regression is absolute to the separate variety set.(Niveditha, Abarna, and Akshaya 2019). Random forest algorithms offer good accuracy and perform faster prediction compared to Logistic regression(Dileep, Navaneeth, and Abhishek 2021)relies on the cloth algorithm program and uses ensemble learning techniques.It creates as several trees on the set of the info and combines the output of all the trees. During this manner it reduces over lifting downside in call trees and additionally reduces the variance and so improves the accuracy.

CONCLUSION

The outcome of the Random forest classifier 97.12% has better accuracy than logistic regression 95.34% in predicting credit card fraud detection .It would be feasible to work on Random forest classifier than Logistic regression for credit card fraud detection.

DECLARATION(S)

Conflicts of Interest

No conflict of interests in this manuscript

Authors Contributions

Author KL was introduced in data collection, data analysis, and Manuscript writing and Author SSA was involved in conceptualization, data validation, and critical review 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 enabling us to complete the study.

1.Stigmata Techno Solutions.

2.Saveetha University.

3. Saveetha Institute of Medical and Technical Sciences.
4. Saveetha School of Engineering

REFERENCES

1. Alenzi, Hala Z., and O. Nojood. 2020. "Fraud Detection in Credit Cards Using Logistic Regression." *International Journal of Advanced Computer Science and Applications*. https://doi.org/10.14569/ijacsa.2020.0111265.
2. Bahnsen, Alejandro Correa, Aleksandar Stojanovic, Djamila Aouada, and Bjorn Ottersten. 2013. "Cost Sensitive Credit Card Fraud Detection Using Bayes Minimum Risk." *2013 12th International Conference on Machine Learning and Applications*. https://doi.org/10.1109/icmla.2013.68.
3. Baskar, M., R. Renuka Devi, J. Ramkumar, P. Kalyanasundaram, M. Suchithra, and B. Amutha. 2021. "Region Centric Minutiae Propagation Measure Orient Forgery Detection with Finger Print Analysis in Health Care Systems." *Neural Processing Letters*, January. https://doi.org/10.1007/s11063-020-10407-4.
4. Bhanu Teja, N., Yuvarajan Devarajan, Ruby Mishra, S. Sivasaravanan, and D. Thanikaivel Murugan. 2021. "Detailed Analysis on Sterculia Foetida Kernel Oil as Renewable Fuel in Compression Ignition Engine." *Biomass Conversion and Biorefinery*, February. https://doi.org/10.1007/s13399-021-01328-w.
5. Bhavikatti, Shaeesta Khaleelahmed, Mohmed Isaqali Karobari, Siti Lailatul Akmar Zainuddin, Anand Marya, Sameer J. Nadaf, Vijay J. Sawant, Sandeep B. Patil, Adith Venugopal, Pietro Messina, and Giuseppe Alessandro Scardina. 2021. "Investigating the Antioxidant and Cytocompatibility of Mimusops Elengi Linn Extract over Human Gingival Fibroblast Cells." *International Journal of Environmental Research and Public Health* 18 (13). https://doi.org/10.3390/ijerph18137162.
6. 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.
7. 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.
8. Hussein, Ameer Saleh, Rihab Salah Khairy, Shaima Miqdad Mohamed Najeeb, and Haider Th Salim Alrikabi. 2021. "Credit Card Fraud Detection Using Fuzzy Rough Nearest Neighbor and Sequential Minimal Optimization with Logistic Regression." *International Journal of Interactive Mobile Technologies (iJIM)*. https://doi.org/10.3991/ijim.v15i05.17173.
9. Karobari, Mohmed Isaqali, Syed Nahid Basheer, Fazlur Rahman Sayed, Sufiyan Shaikh, Muhammad Atif Saleem Agwan, Anand Marya, Pietro Messina, and Giuseppe Alessandro Scardina. 2021. "An In Vitro Stereomicroscopic Evaluation of Bioactivity between Neo MTA Plus, Pro Root MTA, BIODENTINE & Glass Ionomer Cement Using Dye Penetration Method." *Materials* 14 (12). https://doi.org/10.3390/ma14123159.
10. Karthigadevi, Guruviah,

Sivasubramanian Manikandan, Natchimuthu Karmegam, Ramasamy Subbaiya, Sivasankaran Chozhavendhan, Balasubramani Ravindran, Soon Woong Chang, and Mukesh Kumar Awasthi. 2021. "Chemico-Nanotreatment Methods for the Removal of Persistent Organic Pollutants and Xenobiotics in Water - A Review." *Bioresource Technology* 324 (March): 124678.

11. Kaur, Bhagwant Jot, and Rakesh Kumar. 2020. "A Hybrid Approach for Credit Card Fraud Detection Using Naive Bayes and Voting Classifier." *Proceeding of the International Conference on Computer Networks, Big Data and IoT (ICCBI - 2019)*. https://doi.org/10.1007/978-3-030-43192-1_81.
12. 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.
13. Kulkarni, Abhilasha. 2019. "Credit Card Fraud Detection Using Random Forest and Local Outlier Factor." *International Journal for Research in Applied Science and Engineering Technology*. https://doi.org/10.22214/ijraset.2019.4209.
14. Kumar, M. Suresh, M. Suresh Kumar, V. Soundarya, S. Kavitha, E. S. Keerthika, and E. Aswini. 2019. "Credit Card Fraud Detection Using Random Forest Algorithm." *2019 3rd International Conference on Computing and Communications Technologies (ICCCT)*. https://doi.org/10.1109/iccct2.2019.8824930.
15. Kumar, Tulika. 2021. "Comparison of Logistic Regression and Decision Tree Method for Credit Card Fraud Detection." *International Journal for Research in Applied Science and Engineering Technology*. https://doi.org/10.22214/ijraset.2021.34241.
16. Muthukrishnan, Lakshmipathy. 2021. "Nanotechnology for Cleaner Leather Production: A Review." *Environmental Chemistry Letters* 19 (3): 2527–49.
17. Niveditha, G., K. Abarna, and G. V. Akshaya. 2019. "Credit Card Fraud Detection Using Random Forest Algorithm." *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*. https://doi.org/10.32628/cseit195261.
18. Preethi, K. Auxzilia, K. Auxzilia Preethi, Ganesh Lakshmanan, and Durairaj Sekar. 2021. "Antagomir Technology in the Treatment of Different Types of Cancer." *Epigenomics*. https://doi.org/10.2217/epi-2020-0439.
19. Sawant, Kashmira, Ajinkya M. Pawar, Kulvinder Singh Banga, Ricardo Machado, Mohmed Isaqali Karobari, Anand Marya, Pietro Messina, and Giuseppe Alessandro Scardina. 2021. "Dentinal Microcracks after Root Canal Instrumentation Using Instruments Manufactured with Different NiTi Alloys and the SAF System: A Systematic Review." *NATO Advanced Science Institutes Series E: Applied Sciences* 11 (11): 4984.
20. Shanmugam, Vigneshwaran, Rhoda Afriyie Mensah, Michael Försth, Gabriel Sas, Ágoston Restás, Cyrus Addy, Qiang Xu, et al. 2021. "Circular Economy in Biocomposite Development: State-of-the-Art, Challenges and Emerging Trends." *Composites Part C: Open Access* 5 (July): 100138.
21. 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.

22. Veerasimman, Arumugaprabu, Vigneshwaran Shanmugam, Sundarakannan Rajendran, Deepak Joel Johnson, Ajith Subbiah, John Koilpichai, and Uthayakumar Marimuthu. 2021. "Thermal Properties of Natural Fiber Sisal Based Hybrid Composites – A Brief Review." *Journal of Natural Fibers*, January, 1–11.
23. V, Gokula Krishnan, and Krishnan V. Gokula. 2019. "Credit Card Fraud Detection Using Random Forest Algorithm." *International Journal for Research in Applied Science and Engineering Technology*. https://doi.org/10.22214/ijraset.2019.3215.
24. 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.
25. Zhang, Dongfang, Basu Bhandari, and Dennis Black. 2020. "Credit Card Fraud Detection Using Weighted Support Vector Machine." *Applied Mathematics*. https://doi.org/10.4236/am.2020.111208 7.

TABLES AND FIGURES

Table 1. Comparison of prediction accuracy using Random forest classifier and Logistic regression.The accuracy of Random forest classifier is 97% and the accuracy of Logistic regression is 95%.

Execution	Random Forest Classifier	Logistic Regression
1	97.12	95.34
2	97.0	95.76
3	97.3	95.45
4	97.9	97.45
5	97.51	97.23
6	97.63	98.65
7	97.77	98.23
8	97.23	98.12
9	97.85	98.34
10	97.34	98.34

Table 2. Mean of the Random Forest Classifier is 97.1460 and Logistic Regression is 95.9190 and standard deviation of Random Forest Classifier is 6.1217 and Logistic Regression is 5.7113 and std.error of Random Forest Classifier is 1.9359 and Logistic Regression is .18061.

Groups	N	Mean	Std.deviation	Std.error mean
Accuracy RFC	10	97.1470	.61217	.19359
LR	10	95.9190	.57113	.18061
Errors RFC	10	1.8530	.61217	.19359
LR	10	3.1110	.58005	.18343

Table 3. Independent sample Test the accuracy increases as the error rate decreases. The 2-tailed Significance is 0.001($p < 0.05$)

Accuracy	Levene's Test for Equality of Variances		T-test of Equality of Means					95% of the confidence interval of the Difference	
	F	Sig.	t	df	Sig (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Equal Variance Assumed	.127	0.02	4.638	18	0.001	1.22800	.26475	.67177	1.78423
Equal Variance Not Assumed	.042	0.02	-4.717	18	0.001	-1.25800	.2669	-1.81829	-69771

GRAPH:

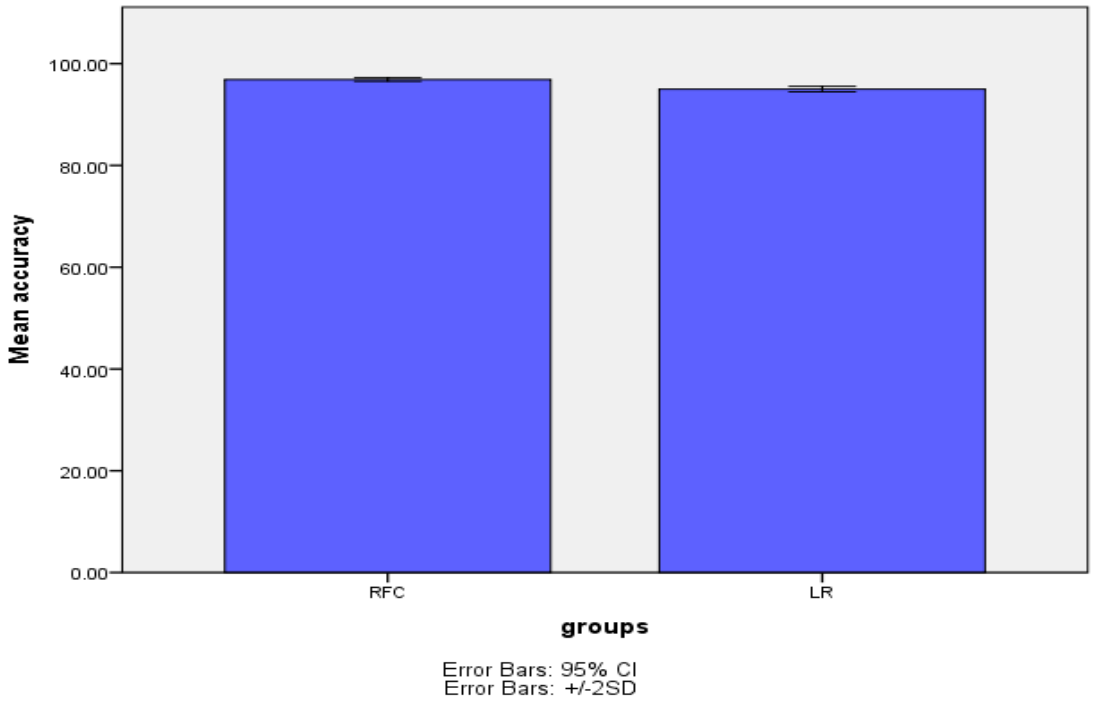

Fig.1. Prediction accuracy for the two algorithms the accuracy of the Random Forest Classifier is better than that of the Logistic Regression X-axis Random Forest Classifier vs Logistic Regression Y-axis mean accuracy of detection +- 2SD.