A Machine Learning Method for Predicting Loan Approval by Comparing the Random Forest and Decision Tree Algorithms.

P. Bhargav[1], K. Sashirekha[2*]

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

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

ABSTRACT

Aim: The objective of this work is to determine an approach in machine learning for loan approval prediction by comparing Random Forest algorithms with Decision Trees. To achieve accuracy novel random forest classifiers are used. **Materials and Methods**: Loan prediction datasets from the kaggle library are used to test accuracy and loss.The total sample size is 20. The two groups considered were Random Forest (N=10) and Decision tree (N=10). The computation is performed using G-power as 80%. **Results:** While the random forest method has a precision of 79.4490% and loss is 21.0310%, a method that looks superior to the traditional decision tree of 67.2860% and loss is 32.7140% respectively. Finally, it seems that the Random Forest method outperforms the Decision tree. RF and DT, The independent sample T-test result of p=0.33 (p>0.05) shows a statistically significant agreement between the two most extensively used machine learning techniques shows that two groups are statistically insignificant with confidence level of 95%. **Conclusion**:Random Forests seem to be more accurate in predicting loan acceptance than Decision Trees.

Keywords: Loan Prediction, Loan Risk, Machine Learning, Decision Tree, Novel Random Forest Classifier, Accuracy.

INTRODUCTION

Loan prediction is a technology that offers you a loan approval interface for the applicant's loan application. Applicants supply personal information to a system, which then determines whether or not a loan is available based on that information (Choden and Unhapipat 2019). The majority of the bank's assets were derived directly from profits gained through loans disbursed by banks. Banks' principal purpose is to invest their money in safe places. While many banks and financial institutions currently give loans after extensive verification and validation, there is no guarantee that the party's nominee would be the most deserving of all applicants (Gramespacher and Posth 2021). In this approach, machine learning is used to predict if an application is secure, and the whole feature validation process is automated (Sheikh, Goel, and Kumar 2020). The size of the loan is also a factor that should be considered when approving the risk of the loan. The chances of getting a small loan approved are higher. In general, a smaller monthly payment increases your chances of getting your loan authorised(Chen, Zhang, and Ng 2018).

The impact of loan predictions are discussed in different forms compared to this paper analysis. A total of 15 publications in IEEE Explore and 7 in

Google Scholar have been published on the topic of loan forecasting. Understand lending processes in order to develop a model for predicting loan risk based on demographic information and other criteria that come together into a more complex approval process, and then to implement this model on cloud-based platforms (Ramachandra et al. 2021). In general, banks rely heavily on loan income to fund their operations. Urbanization has resulted in a huge increase in the number of people seeking loans. It gets more difficult to find someone who is eligible for a loan because of this (Gopinath, Srinivas Shankar Maheep, and Sethuraman 2021). Fraud detection in online banking transactions relied on DTs, deep learning, and Artificial Neural Networks (ANN). To identify counterfeit banknotes, DT, ANN, and SVM algorithms were coupled in another application. When using DT models to produce rules, it was possible to distinguish between real and fake banknotes (Meshref 2020). Decisions on whether or not an application is accepted or rejected are made using data from prior candidates. As many individuals seek bank loans each day, yet the bank's reserves are limited, it is difficult to meet demand. To create an accurate forecast in this situation, certain lessons would be quite helpful. -function (Singh et al. 2021). Machine learning is used to build predictive and probabilistic techniques for a particular problem in loan prediction in this study. If a loan is authorised for the collection of data about an application, this study employs logistic regression to determine this (Vaidya 2017). Several types of loan monitoring are used to examine the problem of trust among loan officers while making approval judgements. Loan officers were requested to assess probability of loan acceptance based on a standard loan monitoring or continuous reporting capacity (Searcy and Ward 2011).

(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)

The research gap identified it is less accurate than the Decision Tree algorithm. While examining and adding human input to the dataset, the proportion of loan acceptance predictions is revealed to be quite low. The present model's categorization of loan approval prediction is less accurate since it uses limited functions. The ultimate goal of research is to increase the precision of determining loan approval prediction, and to reduce loss of data while training and testing dataset. The novel Random Forest Classifier used to achieve accuracy.

MATERIALS AND METHODS

The Data Analytics Lab at Saveetha School of Engineering is the site of the planned research. The Random Forest method and the Decision Tree algorithm were categorised into two categories for this investigation. Each group has ten participants as its sample size. G-power is 80 percent, with a 95 percent confidence range, Alpha and Beta are each set at 0.05 and 0.2 (Ambika, Ambika, and Biradar 2021).

The datasets are downloaded from kaggle and named as loan approval prediction dataset. The extracted data sets contain information about Variables and Description with unique features like

Loan_ID, Gender, Married, Dependents.Both the training and the testing portions of the dataset were separated. 70% of the information was used for training, while the remaining 30% was used for testing.. The technique was developed by comparing test and training datasets and then implementing it. For exhibiting this research work, a jupyter notebook is used along with a laptop with AMD ryzen 5 processor, 8GB RAM, along with 512 SSD with 64 bit operating system.

RANDOM FOREST ALGORITHM

Supervised algorithm learning is the backdrop behind the Random Forest algorithm. It is used as part of a larger ensemble. Using a variety of basic models, it aims to provide the most accurate forecast of loan default risk possible. There are several decision trees in a Random Forest states classifier, and the mean accuracy of these decision trees is used to improve accuracy in the dataset as a whole. Random Forest, as opposed to depending on a decision tree, takes the forecasts from each tree and predicts the final outcome based on the majority vote (Arutjothi and Senthamarai 2017).Table 1 describes the complete pseudo code for Random Forest Algorithm.

DECISION TREE

The supervised learning family includes Decision Tree. It is mostly utilised for loan classification and prediction, as well as problem-solving techniques. To set a high precision, tree-based learning algorithms are extensively utilised with prediction models that use supervised learning approaches. Both category and digital data can be handled by decision trees. Decision trees can help you reach a conclusion or make a decision.Table 2 explains pseudocode for DT supervised algorithms. Table 5 represents accuracy of loan approval prediction of classification using Decision Tree (Tchakoute-Tchuigoua and Soumaré 2019).Table 2 describes pseudo code for Decision Tree Algorithm.

Equation 1 determines the level of precision.

Accuracy = TP+TN/ TP+TN+FP+FN (1)

Where,

The TP represents the model's definition of "true positives," which is the number of such results.

The FP is a measure of how many false positives the model picked up.

TN is a metric for expressing the total number of actual negatives in a given model.

The amount of false - negative in the data set is known as the model-specific FN.

Statistical Analysis

IBM SPSS is utilised here for statistical implementation V22.0. Statistical Package for Social Sciences (SPSS) Mean, standard error, and graphs are all calculated and plotted using this software. The independent variables are Variable Name, Description, Type. The dependent variable is ‘accuracy’. For each group, a sample size of 10 is used, and accuracy is tested as a separate testing variable ((Ambika, Ambika, and Biradar 2021) in an independent samples-t-test.

RESULTS

The experimental results are carried out on Random Forest algorithm and Decision Tree algorithm where performance is measured based on accuracy. Table 3 describes Datasets for loan approval prediction. The classification accuracy and corresponding loss obtained with both the classifiers in the number of test runs shown in Table 4 and Table 5. Random Forest and Decision Tree algorithms are compared in Table 6. While Decision Tree's accuracy is 67.2860%, Random Forest's is 70.94490 % accurate.

Table 7 represents the independent sample test that has for determining constant variance expected and constant variance not assumed by the Random Forests and Tree Based algorithm and it also shows mean difference, standard error differences with a confidence level of 95%. Independent sample T-test value $p=0.33$ ($p>0.05$) shows that two groups are statistically insignificant

Figure 1 shows a simple bar means graph of accuracy by a group of Random Forest algorithm and Decision Tree algorithm. It is observed that a novel Random Forest Classifier algorithm has a higher significance when compared to Decision Tree algorithm. For Random Forests, error bars are given, and their error rate is lower than for Decision Trees.

DISCUSSION

The Random Forest method was found to have a 79.4490 percent success rate in this investigation.When compared to the Decision Tree, this technique has a higher level of accuracy.

To support this research work,The Random Forest algorithm gave more accurate results, as shown (Calcagnini et al. 2018). There is a considerable improvement in client satisfaction and a reduction in operating expenses (“Videotex Loan Approval Service” 1984). However, the bank can only reap the benefits if it has a reliable model for predicting which customers' loans it should accept and which it should refuse in order to reduce the chance of loan default (Turiel and Aste 2020). To oppose this research work, The majority of diagrams depict the nature of a loan application from the best available inspection it is set up for long haul credit and momentary advance. The predictive model aids in the analysis of various loan limitations (S.m., Karthikeyan, and Ravikumar 2021). In spite of this, It is assumed that the references cited address the principal hypothetical difficulties and provide access to key portions of the writing managing such tactics, as well as experts in exciting research headings (Karthiban, Ambika, and Kannammal 2019). In the bank's quarterly financial statements, the revenue and profitability are directly impacted by the acceptance or refusal of each loan application. Loan approval is critical, but getting to that point is a complicated procedure full of uncertainties(Alaradi and Hilal 2020).

When compared to earlier studies, the Random Forest method performed better. The limitation of this proposed model is determining the loan approval prediction is mainly classified using only a smaller number of attributes. Predicting loan amounts is beneficial to both bank workers and applicants. The loan risk system automatically determines the weight of each loan processing characteristic, and the same features are processed on fresh test data. In the future, this paper work might be expanded to a more advanced level.

CONCLUSION

This research work indicates that the Random Forest algorithm based model for detection of loan approval using Random Forest model performs better than Decision Tree with improved accuracy of 79.4490%.

DECLARATION

Conflict of Interests

The author declares no conflict of interest.

Authors Contribution

Author PB was involved in the data collection, data analysis, Manuscript writing. Author PSP assisted in conceptualization, data validation and critical review of the manuscript.

Acknowledgments

The author wishes to express his gratitude to the Saveetha School of Engineering, Saveetha Institute of Medical and Technical Science, Saveetha University for providing the opportunities and facilities needed to carry out research studies.

Funding

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

1. Insysness Technology Pvt. Ltd. Chennai.
2. Saveetha University.
3. Saveetha Institute of Medical and Technical Sciences.
4. Saveetha School of Engineering.

REFERENCES

1. Alaradi, Mohamed, and Sawsan Hilal. 2020. "Tree-Based Methods for Loan Approval." *2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI)*. https://doi.org/10.1109/icdabi51230.2020.9325614.
2. Ambika, Ambika, and Santosh Biradar. 2021. "Survey on Prediction of Loan Approval Using Machine Learning Techniques." *International Journal of Advanced Research in Science, Communication and Technology*. https://doi.org/10.48175/ijarsct-1165.
3. Arutjothi, G., and C. Senthamarai. 2017. "Prediction of Loan Status in Commercial Bank Using Machine Learning Classifier." *2017 International Conference on Intelligent Sustainable Systems (ICISS)*. https://doi.org/10.1109/iss1.2017.8389442.
4. 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.
5. 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.
6. 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.

7. Calcagnini, Giorgio, Rebel Cole, Germana Giombini, and Gloria Grandicelli. 2018. "Hierarchy of Bank Loan Approval and Loan Performance." *Economia Politica*. https://doi.org/10.1007/s40888-018-0109-3.
8. Chen, Ya-Qi, Jianjun Zhang, and Wing W. Y. Ng. 2018. "Loan Default Prediction Using Diversified Sensitivity Undersampling." *2018 International Conference on Machine Learning and Cybernetics (ICMLC)*. https://doi.org/10.1109/icmlc.2018.8526936.
9. Choden, Sonam, and Suntaree Unhapipat. 2019. "Statistical Model for Personal Loan Prediction in Bhutan." *Journal of Advanced Research in Dynamical and Control Systems*. https://doi.org/10.5373/jardcs/v11/20192587.
10. Gopinath, Mahankali, K. Srinivas Shankar Maheep, and R. Sethuraman. 2021. "Customer Loan Approval Prediction Using Logistic Regression." *Advances in Parallel Computing*. https://doi.org/10.3233/apc210103.
11. Gramespacher, Thomas, and Jan-Alexander Posth. 2021. "Employing Explainable AI to Optimize the Return Target Function of a Loan Portfolio." *Frontiers in Artificial Intelligence* 4 (June): 693022.
12. 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.
13. Karthiban, R., M. Ambika, and K. E. Kannammal. 2019. "A Review on Machine Learning Classification Technique for Bank Loan Approval." *2019 International Conference on Computer Communication and Informatics (ICCCI)*. https://doi.org/10.1109/iccci.2019.8822014.
14. Karthigadevi, Guruviah, Sivasubramanian Manikandan, NatchimuthuKarmegam, Ramasamy Subbaiya, SivasankaranChozhavendhan, 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.
15. Meshref, Hossam. 2020. "Predicting Loan Approval of Bank Direct Marketing Data Using Ensemble Machine Learning Algorithms." *International Journal of Circuits, Systems and Signal Processing*. https://doi.org/10.46300/9106.2020.14.117.
16. Muthukrishnan, Lakshmipathy. 2021. "Nanotechnology for Cleaner Leather Production: A Review." *Environmental Chemistry Letters* 19 (3): 2527–49.

17. 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.
18. Ramachandra, H. V., G. Balaraju, R. Divyashree, and Harish Patil. 2021. "Design and Simulation of Loan Approval Prediction Model Using AWS Platform." *2021 International Conference on Emerging Smart Computing and Informatics (ESCI)*. https://doi.org/10.1109/esci50559.2021.9397049.
19. Sawant, Kashmira, Ajinkya M. Pawar, Kulvinder Singh Banga, Ricardo Machado, Mohmed IsaqaliKarobari, 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. Searcy, Dewayne L., and Terry J. Ward. 2011. "Loan Officer Confidence, Continuous Reporting, And The Loan Approval Process." *Journal of Applied Business Research (JABR)*. https://doi.org/10.19030/jabr.v25i5.1005.
21. Shanmugam, Vigneshwaran, Rhoda Afriyie Mensah, Michael Försth, Gabriel Sas, ÁgostonRestás, 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.
22. Sheikh, Mohammad Ahmad, Amit Kumar Goel, and Tapas Kumar. 2020. "An Approach for Prediction of Loan Approval Using Machine Learning Algorithm." *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*. https://doi.org/10.1109/icesc48915.2020.9155614.
23. Singh, Vishal, Ayushman Yadav, Rajat Awasthi, and Guide N. Partheeban. 2021. "Prediction of Modernized Loan Approval System Based on Machine Learning Approach." *2021 International Conference on Intelligent Technologies (CONIT)*. https://doi.org/10.1109/conit51480.2021.9498475.
24. S.m., Karthikeyan, S. M. Karthikeyan, and Pushpa Ravikumar. 2021. "A Comparative Analysis of Feature Selection for Loan Prediction Model." *International Journal of Computer Applications*. https://doi.org/10.5120/ijca2021920992.
25. Tchakoute-Tchuigoua, Hubert, and Issouf Soumaré. 2019. "The Effect of Loan Approval Decentralization on Microfinance Institutions' Outreach and Loan Portfolio Quality." *Journal of Business Research*. https://doi.org/10.1016/j.jbusres.2018.09.021.
26. Turiel, J. D., and T. Aste. 2020. "Peer-to-Peer Loan Acceptance and Default Prediction with Artificial Intelligence." *Royal Society Open Science* 7 (6): 191649.
27. Vaidya, Ashlesha. 2017. "Predictive and Probabilistic Approach Using Logistic Regression: Application to

Prediction of Loan Approval.” *2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*. https://doi.org/10.1109/icccnt.2017.8203946.

28. Veerasimman, Arumugaprabu, Vigneshwaran Shanmugam, Sundarakannan Rajendran, Deepak Joel Johnson, Ajith Subbiah, John Koilpichai, and UthayakumarMarimuthu. 2021. “Thermal Properties of Natural Fiber Sisal Based Hybrid Composites – A Brief Review.” *Journal of Natural Fibers*, January, 1–11.
29. “Videotex Loan Approval Service.” 1984. *Computer Communications*. https://doi.org/10.1016/0140-3664(84)90055-0.

TABLES AND FIGURES

Table 1. Pseudo code for Random Forest Algorithm

S.No	**Steps**
1	Input all required packages
2	Load data into program using Pandas Library
3	Duplicate values,Stop words in dataset are removed
4	Word count package is imported and count repetition of words
5	Prepare model and train data
6	Import the Machine learning algorithms and predict the output
7	OUTPUT: loan approval prediction as an output.

Table 2. Pseudo code for Decision Tree Algorithm

S.No	**Steps**
1	Assign a variable name for respective datasets
2	Using pandas library complete the process of Data Cleaning
3	Extract common words by using word cloud
4	Import scikit learn library
5	Import Decision Tree algorithm from Scikit learn
6	Use the pipeline() command and Train model with 80% of the data
7	Predict the output

Table 3. Datasets for loan approval prediction

Variable Name	**Description**	**Type**
Loan_ID	Unique Loan ID	Integer
Gender	Male/Female	Character
Maritial_Status	Applicant married(Y/N)	Character
Department	Number of dependents	Integer
Education_Qualification	Graduate/Undergraduate	String
Self_Implyed	Self Employed(Y/N)	Character
Applicant_Income	Applicant income	Integer
Co_Applicant_Income	Co Applicant income	Integer
Loan_Amount	Loan amount in thousands	Integer
Loan_Status	Loan Approved(Y/N)	Character

Table 4. Accuracy of loan approval prediction classification using Random Forest algorithm (Mean Accuracy=79.4490%, mean Loss=21.0310%)

Test	**Accuracy**	**Loss**
Test 1	81.20	18.80
Test 2	80.90	19.10
Test 3	81.10	18.90
Test 4	79.20	20.80
Test 5	79.95	20.05
Test 6	79.05	21.95
Test 7	79.35	21.65
Test 8	78.12	22.78
Test 9	78.50	22.50
Test 10	77.12	23.78

Table 5. Accuracy of loan approval prediction classification using Decision Tree algorithm (Mean Accuracy=67.2860%, mean Loss=32.7140%)

Test	Accuracy	Loss
Test 1	67.47	32.53
Test 2	67.35	32.65
Test 3	68.12	31.88
Test 4	67.35	32.65
Test 5	67.95	32.05
Test 6	66.85	33.15
Test 7	67.12	32.88
Test 8	66.75	33.25
Test 9	67.80	32.20
Test 10	66.10	33.90

Table 6. Group Statistics Random forest and Decision algorithm with the mean value of 79.45% and 67.28%.

	GROUPS	N	MEAN	std.Deviation	std.Error Mean
ACCURACY	RF	10	79.4490	1.35279	0.42779
	DT	10	67.2860	0.61049	0.19305
LOSS	RF	10	21.0310	1.77215	0.56040
	DT	10	32.7140	0.61049	0.19305

Table 7. Independent Samples T-test shows significance value achieved is p=0.33 (p>0.05), which shows that two groups are statistically insignificant.

		F	Sig	t	df	sig(2-trailed)	Mean difference	Std Error difference	Lower	Upper

ACCURACY	Equal variance assumed	5.333	0.33	25.916	18	<0.001	12.16300	0.46933	11.17697	13.14903
	Equal variance not assumed			25.916	12.520	<0.001	12.16300	0.46933	11.14510	13.18090
LOSS	Equal variance assumed	13.947	0.002	-19.711	18	<0.001	-11.68300	0.59272	-12.98605	-10.43773
	Equal variance not assumed			-19.711	11.106	<0.001	-11.68300	0.59272	-12.98605	-10.37995

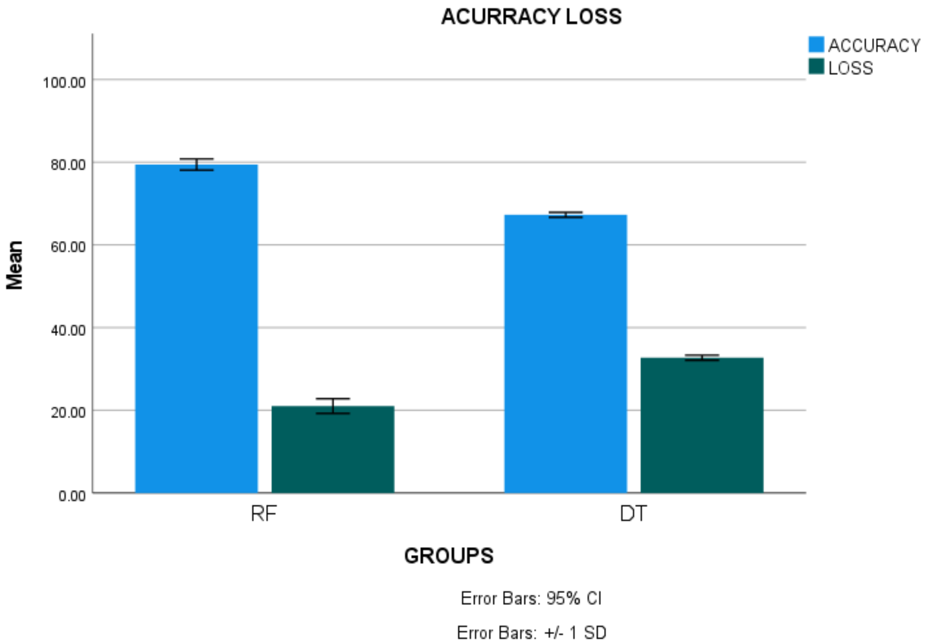

Fig. 1. The mean accuracy of the algorithms Random Forest and Decision Tree is compared. Random Forest has a better mean accuracy and a slightly better standard deviation than Decision Tree. X Axis: Random Forest vs Decision Tree. Y Axis : Mean Accuracy of detection = +/- 1 SD.