Title Page:

Prediction of credit card approval using Random Forest Algorithm over Support vector machine

Seeram Balu[1] , Dr.K.Somasundaram[2]

Seeram Balu[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, Pin: 602105.
seerambalu18@gmail.com

Dr.K.Somasundaram[2]
Project Guide, Corresponding Author,
Department of Computer Science And Engineering,
Saveetha School of Engineering ,
Saveetha Institute of Medical And Technical Sciences,
Saveetha University, Chennai, Tamil Nadu, India, Pin :602105.
subbiahs.sse@saveetha.com

Keywords: Credit Card approval, Random Forest Algorithm, Support vector machine, Data Manipulation, Hyper parameters,*Reduction*.

ABSTRACT

Aim: This study aims to develop a model that can forecast if a financial institution will be able to approve credit cards for its clients. In order to prevent fraud, which can result in financial institutions losing money, this model can help an organisation come to a clear conclusion about whether to accept or reject a card. **Materials and methods:** Classification is performed by Randomforest algorithm (n=10) over Support vector machine (n=10) is for false rate detection. The statistical test difference between g-power 0.08 and alpha value is α=0.05. **Result and Discussion:** The analysis of the result shows that the Random forest algorithm has a high accuracy compared to Support vector machine. The mean value is 78.5450 and mean accuracy of detection is +/-1SD and significant value is p=0.0.001, ($p<0.05$) from independent sample T size. This indicates that there is a statistically significant difference between the two algorithms. **Conclusion:** The model is used to forecast and analyse whether a financial institution will provide a credit card or not to its customers. Our research reveals that prior default, years worked, credit score, and debt are the four most important factors that any financial business host would take into account when deciding whether to give a credit card.

Keywords: Credit Card approval, Random Forest Algorithm, Support vector machine, Data Manipulation , Hyper parameters, *Reduction*.

INTRODUCTION

The board of credit risk in banks is tasked with determining the likelihood of a client defaulting or experiencing credit decay, as well as the cost of doing so if it occurs(John, Tullu, and Gupta 2022). It is critical to analyse main elements and anticipate the likelihood of consumers defaulting based on their circumstances. This is where a machine learning model comes in helpful, allowing banks and other large financial institutions to forecast whether or not the person to whom they are lending will default(Gwadz et al. 2022). Using Python, this project creates a machine learning model with the highest accuracy achievable. We begin by loading and viewing the dataset.A large number of credit card applications are received by banks(Olaniran et al. 2022)(Subramanian et al. 2022). Many of the application reduction are turned down for a number of reasons, such as high loan balances or low income(Financial Crisis Inquiry Commission 2011). A reduction in credit score is done if the client is overdue on their loan repayment.

In the last four years, 17,400 articles on credit card approval using machine learning have been published in Google Scholar, with 1,773 articles available in ScienceDirect. One of the most well-known options is to use a credit card. Because credit cards are a simple way to pay, the majority of consumers utilise them to perform their transactions(Bartram, Branke, and Motahari 2020). Numerous financial institutions, including national and commercial banks, make decisions based on consumer data such as fundamental information, lifestyles, compensation, annual and monthly returns, and current source of livelihood income.(United Nations 2020).

Before an application is considered, all of this information is analysed.(Majid et al. 2022)(Chisholm et al. 2022).A reduction in credit score is done if the client is unwilling to repay their loan repayment. In order to maintain a low credit utilization rate, consider reducition your spending or making periodic bill payments throughout your billing cycle.

The objective of a credit scoring model is to segregate credit applicants into two categories: those with "great credit," who are more prone to paying their bills, and those with "poor credit," who should be discard credit since they are more likely to miss payments. The aim of the study is to expand the input dataset size while enhancing algorithm performance.. Previous research had oversampled their datasets. The previous study's input data set was made up of incorrect datasets(Vajjala et al. 2020)(Brownlee 2020). Positive classes get a lot of attention. The number of tuples obtained after cleaning the dataset is quite little(Posen and Changyong 2013). As a result, the input dataset is exceedingly small. The precision of the current study is really low(Fernández et al. 2018).

MATERIALS AND METHODS

SIMATS Saveetha School of Engineering's computer science and engineering department carried out the study. Each group was given a total of n=10 iterations in order to increase accuracy. The data set taken from the kaggle website credit card approval dataset. With 95% confidence and 80% pretest power, the experimental arrangement is maintained.

The dataset which proposed work used in this paper is credit card approval prediction dataset. The Dataset was collected from the open source Kaggle platform. The Hardware configuration were HP i5 processor with a RAM size of 12GB was used.The system type used was 64-bit, OS, x64 based processor with HDD of 917 GB. Windows was used as the operating system, and the Python programming language was employed with the JupyterLab tool.

Random forest

The number of trees in the forest was originally set to 1000, and the pseudo random number generator's random seed was set to 40(Zhao 2013). There are two sections to the data set: training (80%) and testing (20%) (20 percent)(Reinert et al. 2022). (Less than 20% of the total). Before making data into two sections, data manipulation is an important task. Data Manipulation deals with missing values, string data, which parameters should be considered etc.Reduction of parameters is not allowed. By reduction the algorithm will become overrated or underrated.It selects samples at random, and decision trees are built for each sample to predict the outcome(Milinkovic and Malesevic 2012). Every possible option was voted on, and the outcome that received the most votes was chosen as the final result(Xu, Tong, and Meng 2016). In Random forest dealing with hyper parameters is tricky because of hyper parameters the model

can become overrated or underrated. Hyper parameters are maximum depth of tree , number of decision trees, minimum number of samples required to split etc. After setting Hyper parameters check the model whether it is overrated or underrated by using testing data or new data.

Pseudocode for Random Forest Algorithm

Input: Dataset for predicting credit card approval.
Step-1: Import and read the dataset.
Step-2: Choose the features at random from the dataset.
Step-2: Data manipulation
Step-5: Dealing with hyper parameters
Step-4:Create the RF classifier model using valid hyper parameters .
Step-6:Train the random forest regression model using the train dataset.
Step-7:Calculate the error between expected output and actual output .
Step-9: Check whether the model is underrated or overrated
Output: Accuracy in %

Support vector Machine

In a high-dimensional setting, the SVM's primary objective is to determine an ideal hyperplane for a variety of scenarios(Schalley and Springer 2009). To construct this model, you'll need more than one hyperplane. This method employs the blister vector, which contains the information that is closest to the closed surface and coordinates with the best choice surface. It divides data into categories by projecting input vectors into a high-dimensional space and creating a hyperplane(Vijayalakshmi and Muruganand 2019). This method is most commonly used to solve quadratic programming and non-convex, unconstrained minimization issues(Schalley and Springer 2009). SVM algorithm convert low-dimensional to high-dimensional but reduction to low-dimensional is not possible. The SVM approach is the most successful in the classification procedure. The tensor transforms a hyperplane from a low-dimensional to a high-dimensional state(Panesar 2019**).** Before fitting into model data manipulation is a necessary step to be followed. In data manipulation missing data is resolved, removes unwanted data, categorical data is encoded.

Pseudocode for support vector algorithm

Input: credit card approval prediction dataset.
Step-1:Declare all the required library files.
Step-2: Import dataset and assign it to a variable.
Step-3:Display the attributes that are present in the dataset.
Step-4: Data manipulation
Step-5:Train and test the Support Vector Regression Model using train dataset and test dataset .

Step-6:Display the accuracy algorithm.
Step-7:Plot the graph for the accuracy obtained.
Output: Accuracy in %

Statistical Analysis

A t-test is a form of presumed steady used to see if there is a significant difference in the means of two groups that are integrated in some way(Privitera 2011). The output for the grouped statistics was obtained using the spss tool (Bhattacherjee 2012). The t-test is one of many procedures used in statistics for speculation testing.The IBM SPSS version was the statistical programme employed in this study (Cooksey 2020). For this study, dependent variables include attributes like update and transaction class. In SPSS, the datasets are prepared using sample size as 10 for the random forest algorithm and support vector machine(Saunders and Allen 2002). The Groupid for the random forest algorithm is 1 and the Groupid for the support vector machine is 2. The Groupid is a grouping variable and the accuracy is a testing variable(DeCoursey 2003).

RESULTS:

This data is used for the analysis of the Random forest algorithm andSupport vector machine algorithm.These ten data samples, together with their loss, are also applied to analyze statistical values that can be used for analogies. In Table 1, it is shown that the accuracy of two algorithms, Random forest algorithm and Support vector machine algorithm for different N values.The group statistics table depicts the number of samples taken, as well as the mean and standard deviation derived for the precision.

Table 3 represents the outcome of the analysis of the Independent samples test which has been performed for the Random Forest algorithm and the Support vector machine algorithm. From Table 3, the significance value for the one tailed test is found to be 0.720, two-tailed is 0.000 and it is found that the Independent samples test has been carried out at Confidence Interval of 95%

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. For the data set, the Independent sample T test is used, with the confidence interval set to 95%. Table 3 shows the independent t sample test calculation for Accuracy and Loss for RF and SVM. Specifies mean difference and standard error difference and the comparative accuracy analysis, mean of loss between the two algorithms are specified. Figure 1 analyzes the mean of accuracy and mean loss among Random forest algorithm and Support vector machine algorithm.

DISCUSSION:

From the given study the accuracy of the Random forest algorithm is 78.5450% when analogized to the accuracy of the Support vector machine is 75.4910%. Sample size is given as 10 analyses of statistics have been done for both the Random forest algorithm and the Support vector machine algorithm in order to compare both the algorithms to find the better analysis algorithm in detection of credit card approval. For the given group, accuracy has been calculated. The mean, standard deviation and the standard error mean values for the Random forest algorithm are 78.5450 1.76550 and .55830 respectively. Similarly for Support vector machine, the mean, standard deviation and the standard error mean values are 75.4910, 1.47678 and .46700 respectively from Table 2.

Compared to previous analysis of Random forest algorithm and Support vector machine algorithm for credit card approval, our research has got better accuracy in detecting the credit card approval analysis. Previously the Random forest algorithm got the accuracy of 75.7% and for the Support vector machine the accuracy was 70.40%. But in our analysis we got the accuracy of the RF is 78.5450% and for the SVM we got the accuracy of 75.4910%. Datasets have been expelled from various resources and these datasets may contain several independent and unwanted attributes are there,this should be removed in order to get the best accuracy(Saravanan. 2022). Hence, the data is tested and trained to get the best output accuracy. Testing takes a long time. This process necessitates extensive training(Dharmendra 2022). Because of the random nature of the dataset, implementation time is likewise lengthy.(Negi et al. 2022). The training data can be changed but it is a time consuming process and if data is trained low, the accuracy will be reduced (Xi et al. 2022).

Testing and preparation of datasets for both the algorithms RF Algorithm and SVM is possible by grouping the datasets. Cleaning process of information can be efficiently increased and less time consuming in execution of the process. The course of time utilization in preparing the dataset can be diminished.

CONCLUSION:

The Random forest(78.5450%) and the Support vector machine(75.4910%) were used to create a model which makes predictions to approve credit cards or not.The Random forest Algorithm approach outperforms the Support vector machine Algorithm method, according to the results of the studies.

DECLARATIONS

Conflict of Interests

No Conflict of Interest in this manuscript.

Authors Contributions

Author SB was involved in data collection, data analysis, and manuscript writing. Author KS was involved in conceptualization, data validation, and critical review of the 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

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

1. Manac Infotech (P) Limited.
2. Saveetha University.
3. Saveetha Institute of Medical and Technical Sciences.
4. Saveetha School of Engineering.

REFERENCES

Bartram, Söhnke M., Jürgen Branke, and Mehrshad Motahari. 2020. *Artificial Intelligence in Asset Management*. CFA Institute Research Foundation.

Brownlee, Jason. 2020. *Imbalanced Classification with Python: Better Metrics, Balance Skewed Classes, Cost-Sensitive Learning*. Machine Learning Mastery.

Chisholm, Leah P., Elisabeth M. Sebesta, Stephanie Gleicher, Melissa Kaufman, Roger R. Dmochowski, and William Stuart Reynolds. 2022. "Urinary Incontinence Product Use and Costs Are Higher in Incontinent Women with Greater Unmet Social Needs." *Neurourology and Urodynamics*, July. https://doi.org/10.1002/nau.25007.

DeCoursey, William. 2003. *Statistics and Probability for Engineering Applications*. Elsevier.

Dharmendra, Dharmish, and M. S. Saravanan. 2022. "Prediction of Heart Failure Using Support Vector Machine Compared with Decision Tree Algorithm for Better Accuracy." *2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS)*. https://doi.org/10.1109/icscds53736.2022.9760989.

Fernández, Alberto, Salvador García, Mikel Galar, Ronaldo C. Prati, Bartosz Krawczyk, and Francisco Herrera. 2018. *Learning from Imbalanced Data Sets*. Springer.

Financial Crisis Inquiry Commission. 2011. *The Financial Crisis Inquiry Report: The Final Report of the National Commission on the Causes of the Financial and Economic Crisis in the United States Including Dissenting Views*. Cosimo, Inc.

Gwadz, Marya, Sabrina R. Cluesman, Robert Freeman, Linda M. Collins, Caroline Dorsen, Robert L. Hawkins, Charles M. Cleland, et al. 2022. *BMC Research Notes* 15 (1): 249.

Milinkovic, Luka, and Branko Malesevic. 2012. "Pseudo-Random Number Generator Analysis Based on the Set of Quadratic Irrationals." *2012 20th Telecommunications Forum (TELFOR)*. https://doi.org/10.1109/telfor.2012.6419266.

Negi, Kanishka, Gaddam Prathik Kumar, Gaurav Raj, Subrata Sahana, and Vishal Jain. 2022. "" "Accuracy of the isolation forest algorithm and local outliers factors in catching credit card fraud. 12th Annual convention of cloud technology, machine learning and Engineering in 2022. Olaniran, Abimbola, Jane Briggs, Ami Pradhan, Erin Bogue, Benjamin Schreiber, Hannah Sarah Dini, Hitesh Hurkchand, Madeleine Ballard. 2022 – . Peterson Institute for International Economics".

Privitera, Gregory J. 2011. *Statistics for the Behavioral Sciences*. SAGE.

Reinert, Tomás, Aline C. Gonçalves, C. A. A. de Resende, and Carlos H. Barrios. 2022. "Implications of the PEARL Trial from the Low- to Middle-Income Countries' Perspectives." *European Journal of Cancer* 173 (July): 30–32.

Saunders, Anthony, and Linda Allen. 2002.

Schalley, Christoph A., and Andreas Springer. 2009. *Mass Spectrometry of Non-Covalent Complexes: Supramolecular Chemistry in the Gas Phase*. John Wiley & Sons.

Subramanian, Sujha, Eleanor Namusoke-Magongo, Patrick Edwards, Millicent Atujuna, Teddy Chimulwa, Dorothy Dow, Emilia Jalil, et al. 2022. "Integrated Health Care Delivery for

Adolescents Living with and at Risk of HIV Infection: A Review of Models and Actions for Implementation." *AIDS and Behavior*, July. https://doi.org/10.1007/s10461-022-03787-2.

United Nations. 2020. *World Economic Situation and Prospects 2020*. United Nations.

Vajjala, Sowmya, Bodhisattwa Majumder, Anuj Gupta, and Harshit Surana. 2020. *Practical Natural Language Processing: A Comprehensive Guide to Building Real-World NLP Systems*. O'Reilly Media.

Vijayalakshmi, S. R., and S. Muruganand. 2019. *Embedded Vision: An Introduction*.

Xi, Yantao, Abdallah M. Mohamed Taha, Anqi Hu, and Xianbin Liu. 2022. "Accuracy Comparison of Various Remote Sensing Data in Lithological Classification Based on Random Forest Algorithm." *Geocarto International*. https://doi.org/10.1080/10106049.2022.2088859.

Xu, Hui, Xiaojun Tong, and Xianwen Meng. 2016. "An Efficient Chaos Pseudo-Random Number Generator Applied to Video Encryption." *Optik*. https://doi.org/10.1016/j.ijleo.2016.07.024.

Zhao, Yangchang. 2013. "Decision Trees and Random Forest." *R and Data Mining*. https://doi.org/10.1016/b978-0-12-396963-7.00004-0.

Bartram, Söhnke M., Jürgen Branke, and Mehrshad Motahari. 2020. Artificial Intelligence in Asset Management. CFA Institute Research Foundation.

Brownlee, Jason. 2020. Imbalanced Classification with Python: Better Metrics, Balance Skewed Classes, Cost-Sensitive Learning. Machine Learning Mastery.

Chisholm, Leah P., Elisabeth M. Sebesta, Stephanie Gleicher, Melissa Kaufman, Roger R. Dmochowski, and William Stuart Reynolds. 2022. "Incontinent women have more unmet social needs use and invest more on incontinence solutions. Diary of Neurourology and Urodynamics https://doi.org/10.1002/nau.25007".

Fernández, Alberto, Salvador García, Mikel Galar, Ronaldo C. Prati, Bartosz Krawczyk, and Francisco Herrera. 2018. Learning from Imbalanced Data Sets. Springer.

Financial Crisis Inquiry Commission. 2011. The Financial Crisis Inquiry Report: The Final Report of the National Commission on the Causes of the Financial and Economic Crisis in the United States Including Dissenting Views. Cosimo, Inc.

Gwadz, Marya, Sabrina R. Cluesman, Robert Freeman, Linda M. Collins, Caroline Dorsen, Robert L. Hawkins, Charles M. Cleland, et al. 2022. "Advancing Behavioral Interventions for African American/Black and Latino Persons Living with HIV Using a New Conceptual Model That Integrates Critical Race Theory, Harm Reduction, and Self-Determination Theory: A Qualitative Exploratory Study." International Journal for Equity in Health 21 (1): 97.

John, Rijo M., Fikru T. Tullu, and Rachita Gupta. 2022. "Price Elasticity and Affordability of Aerated or Sugar-Sweetened Beverages in India: Implications for Taxation." BMC Public Health 22 (1): 1372.

Majid, Hafsa, Lena Jafri, Sibtain Ahmed, Muhammad Abbas Abid, Mohammad Amir, Aamir Ijaz, Aysha Habib Khan, and Imran Siddiqui. 2022. "Publication Dynamics: What Can Be

Done to Eliminate Barriers to Publishing Full Manuscripts by the Postgraduate Trainees of a Low-Middle Income Country?" BMC Research Notes 15 (1): 249.

Milinkovic, Luka, and Branko Malesevic. 2012. "Analysis of a Pseudorandom Number Generator Consisting of a Set of Quadratic Irrational numbers telecommunications forum, held in 2012 (TELFOR). https://doi.org/10.1109/telfor.2012.6419266.

Olaniran, Abimbola, Jane Briggs, Ami Pradhan, Erin Bogue, Benjamin Schreiber, Hannah Sarah Dini, Hitesh Hurkchand, and Madeleine Ballard. 2022. "Stock-Outs of Essential Medicines among Community Health Workers (CHWs) in Low- and Middle-Income Countries (LMICs): A Systematic Literature Review of the Extent, Reasons, and Consequences." Human Resources for Health 20 (1): 58.

Posen, Adam S., and Rhee Changyong. 2013. Responding to Financial Crisis: Lessions From Asia Then, the United States and Europe Now. Peterson Institute for International Economics.

Reinert, Tomás, Aline C. Gonçalves, C. A. A. de Resende, and Carlos H. Barrios. 2022. "Implications of the PEARL Trial from the Low- to Middle-Income Countries' Perspectives." European Journal of Cancer 173 (July): 30–32.

Saunders, Anthony, and Linda Allen. 2002. Credit Risk Measurement: New Approaches to Value at Risk and Other Paradigms. John Wiley & Sons.

Subramanian, Sujha, Eleanor Namusoke-Magongo, Patrick Edwards, Millicent Atujuna, Teddy Chimulwa, Dorothy Dow, Emilia Jalil, et al. 2022. "Integrated Health Care Delivery for Adolescents Living with and at Risk of HIV Infection: A Review of Models and Actions for Implementation." AIDS and Behavior, July. https://doi.org/10.1007/s10461-022-03787-2.

United Nations. 2020. World Economic Situation and Prospects 2020. United Nations.

Vajjala, Sowmya, Bodhisattwa Majumder, Anuj Gupta, and Harshit Surana. 2020. Practical Natural Language Processing: A Comprehensive Guide to Building Real-World NLP Systems. O'Reilly Media.

Tables and Figures

Table 1. Comparison between Random forest algorithm and Support vector machine algorithm with N=10 samples of the dataset with the highest accuracy of respectively 81.71% and 77.56% using the 80% of training and 20% of testing dataset

Size	**Random Forest algorithm accuracy in %**	**Support Vector Machine algorithm accuracy in %**
1	81.71	77.56
2	80.50	77.56
3	79.60	76.54
4	79.01	76.16
5	78.7	75.7
6	78.41	75.19
7	77.80	74.7
8	77.19	74.31
9	76.43	73.89
10	76.10	73.30

Table 2. Mean, Standard Deviation and Standard Error mean for Random forest algorithm and Support vector machine algorithm are given below.

Accuracy	Groups	N	Mean	Std. Deviation	Std. Error Mean
	Random Forest algorithm	10	78.5450	1.76550	.55830
	SVM	10	75.4910	1.47678	.46700

Table 3: We find the mean and variance values by using Levene's test for equality of variance and t-test for equality means. By assuming equal variance and unequal variance values. And accuracy for both algorithms

		F	**sig.**	**t**	**df**	**Sig(2-tailed)**	**Mean difference**	**std.Error difference**	**Lower**	**Upper**
Accuracy	Equal variance assumed	.132	.720	4.196	18	<.001	3.05400	.72787	1.52481	4.58319
	Equal variance not assumed			4.196	17.455	<.001	3.05400	.72787	1.52138	4.58662

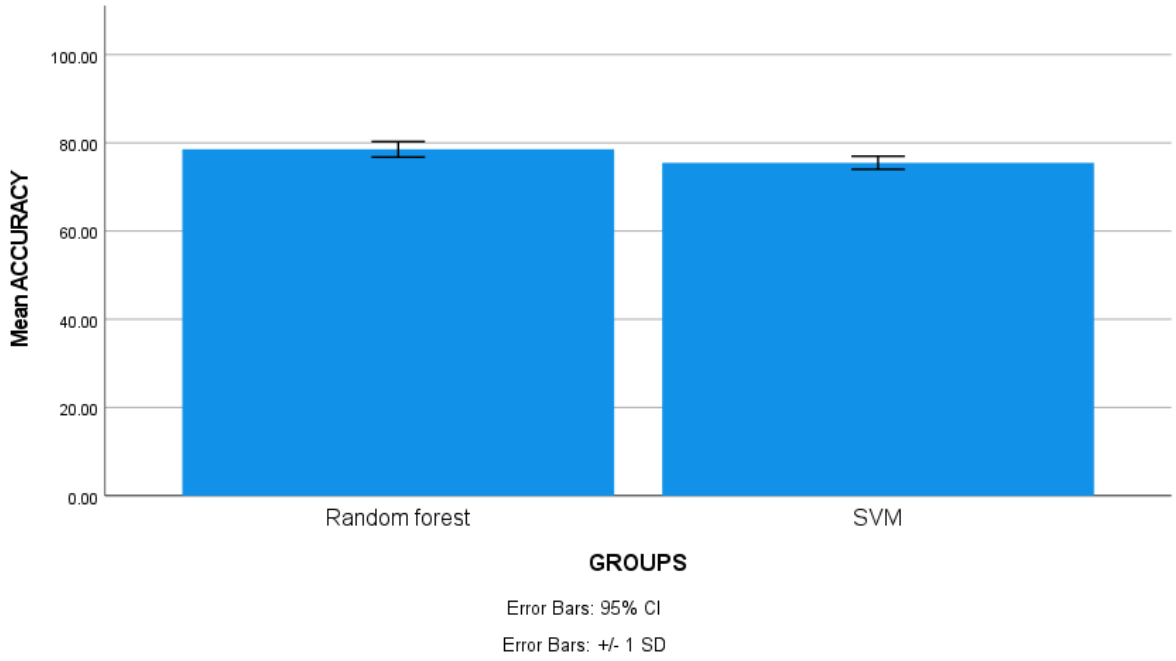

Fig 1. Comparison of Random forest algorithm and Support vector machine algorithm in terms of means accuracy. The mean accuracy of the Random forest algorithm is better than the Support vector machine algorithm . X-Axis: Random forest algorithm Vs :Support vector machine algorithm. Y-Axis:Mean Accuracy of Detection +/-1SD.