Original Article

Classification and Estimation of High-Risk Factors to Low-Risk Factors in Approving Loan through Creditworthiness of Bank Customers using SVM Algorithm and Analyze its Performance over Logistic Regression in terms of Accuracy

Ch.Venkata Sandeep[1], Dr.T. Devi[2]

[1]Research Scholar, Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamilnadu, India, Pincode: 602 105.
[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, Pincode: 602105.

Abstract

Aim: To analyze the accuracy of Support Vector Machine (SVM) algorithms over Logistic Regression (LR) used to approve bank loans. **Materials and Methods**: The existing model uses a Logistic Regression algorithm and the proposed model employs a Novel Support Vector Machine. The 20 sample values are used to find out the mean, std. deviation, std. error means. The sample size was measured as 40 for both the groups using G power (80%). **Results**: The resultant graph explains the comparison of the mean accuracy values of algorithms Novel Support Vector Machine and Logistic Regression where the mean accuracy of the Support Vector Machine is about 69.5% and the mean accuracy value of the Logistic Regression is about 66.5%. The independent sample T-test shows that p=1.0, since p>0.05 there exists insigniface between the SVM and LR. **Conclusion**: The mean accuracy rate of the Novel Support Vector Machine algorithm has been improved to 69.5% compared to Logistic Regression which is having around 66.5% mean accuracy.

Keywords: Logistic Regression, Machine Learning, Novel Support Vector Machine,Creditworthiness, Mean Accuracy, Prediction.

DOI: 10.47750/pnr.2022.13.S04.212

INTRODUCTION

Creditworthiness has tremendously increased in the past decade with an increasing number of transactions happening every day. Typically, creditworthiness assessment involves multiple parameter checks on various levels. Background verification mainly falls (Muflikhah, Hidayat, and Hariyanto 2019) through one of the important categories of evaluation. Background verification of an individual includes checking the payment behavior (Ross Quinlan 1993) patterns and trends, repayment periods, creditworthiness from multiple agencies, etc. It plays a vital role in identifying (Siddig, Ibrahim, and Elkatatny 2021) delinquencies while assessing the applicants for loans. This research highlights the relevance of adopting the best algorithm with the most accuracy for the credit approval procedure by checking credit risks, The examination of either method is (Lee and Verri 2003) performed utilizing real-time customer data gathered from multiple sources. The mean accuracy of the algorithms is used to determine which of the two is the best for the efficient credit application.

There are about 33 articles published since 2017 which are relevant to the topic on IEEE Xplore and about 6 articles on Science Direct.The current existing algorithm, Logistic Regression, a prominent statistical machine learning algorithm, (Nallathambi et al. 2022) is used to classify data using predictive analysis based on the outcome variable of extreme ends, with a logarithmic line separating the ends. Data is taken from Kaggle and utilized for predictions for researching and analyzing creditworthiness. The sigmoid function is employed to construct the model since the output is binary. Multiple characteristics, such as age, credit history, length,

amount, scores, account information, Business Values, Customer Assets, and so on, are used to determine the chance of default per person. The algorithm's primary goal is to guarantee that the data is cleaned to avoid missing values in the numerical data set. Following that, the model begins to be trained using (Bolton 2009) relevant data sets. Because pre-processing is such an important part of the model, it becomes time-consuming and costly. The model also lacks realism since it does not address the issue of class imbalance, and it is hard to analyze non stationary settings. The proposed algorithm, The Novel Support Vector Machine is a supervised machine learning method, used for both regression and classification problems where the latter is focused more. The data points are plotted in (Siddig, Ibrahim, and Elkatatny 2021) the n-dimensional space which is the value of the feature assigned to each coordinate. A hyperplane is then found by performing classification. The kernel trick technique transforms the low dimensional input space into high dimensional space. It essentially converts a non-separable problem into a separable one. It is also memory efficient since it (Deng, Xing, and Cai 2022)uses a subset of training points called support vectors. This type of algorithm works very well when the structure is unknown.

Our institution is passionate about high quality evidence based research and has excelled in various fields (Parakh et al. 2020; Pham et al. 2021; Perumal, Antony, and Muthuramalingam 2021; Sathiyamoorthi et al. 2021; Devarajan et al. 2021; Dhanraj and Rajeshkumar 2021; Uganya, Radhika, and Vijayaraj 2021; Tesfaye Jule et al. 2021; Nandhini, Ezhilarasan, and Rajeshkumar 2020; Kamath et al. 2020).The existing system has issues and major parts (Zhu et al. 2021).It is clear when it is considered the growing demand for new algorithms to address the shortcomings of existing traditional models. The Novel Support Vector Machine method is gaining a competitive advantage over the Logistic Regression technique, which is based on the premise that the dependent and independent variables are always linked. The study's goal is to find the best performing algorithm among the Novel Support Vector Machine and Logistic Regression models for properly analyzing consumers' real-time numerical data. A mean accuracy graph is used for the comparison.

Materials and Methods

The setup of the research has been performed in the Data Analytics Laboratory, Department of CSE in Saveetha School Engineering, Saveetha Institute of Medical and Technical Sciences. The sample size was measured as 40 for both groups using Gpower (80%). The comparison of two algorithms one being Novel Support Vector Machine and the other being Logistic Regression, where a data set named DataSet2 containing 20 rows is used to find the mean accuracy of both algorithms with alpha value 0.05 and beta value is 0.95.

The group 1 is the existing model that uses the Logistic Regression algorithm. The 20 sample values are used to find out the mean, std.deviation, std.error mean. The sigmoid function is calculated using Levene's Test for Equality of Variances both assumed and non-assumed. The Sigmoid 2-tailed function is calculated using the T-test for equality of means.

The group 2 is the proposed Novel Support Vector Machine algorithm that uses 20 sample values where various statistical metrics are evaluated to get to a mean accuracy. This value is used to find the comparison between the existing and proposed models.

This study was implemented using Jupyter lab, and the hardware configuration required is an Intel i3 processor, 50 GB HDD, 4GB RAM, and the software configuration required is a Windows OS. The Study uses a credit card dataset downloaded from Kaggle. The dataset used for the existing model has been imported by Kaggle by downloading the dataset which records around 20 sample values and different attributes related to the output of the data.

Statistical Analysis

The software used to analyze IBM SPSS version 22 (64-bit) is analysis software done by uploading the dataset to the software, which provides output in the form of independent variables N, Mean, Std. Bias, standard error means the mean precision as the output for a given new SVM and logistic regression model. The dependent variables are the output measure in terms of accuracy and cross-validation. The independent variable is the period of time experienced (Zhu 2022).

Results

Table 1 describes the Group 1 Support Vector Machine pseudocode for the factors related to the loan process. It is considered only when the entire process is done.

Table 2 describes the pseudoCode for Logistic Regression which is an optimization algorithm for finding the local minimum of a differentiable function. Gradient descent is used to find values of function parameters.

Table 3 shows the group statistics by comparing its performance for the sample size 20, which yields SVM 69.5% and LR 66.5% of accuracy.

Table 4 shows the independent sample T-test analysis, using the Levene's test for equality variance found the frequency as 0.0, significance as 1.0, at time 2.21 and diff. frequency of 18 for the confidence interval 95%.

Figure 1 shows the graphs of the Input parameters such as the name of the active data set, filters, weight, split files, and the number of rows in the working data file. Missing Value Handling, Syntax, and Resources are described further. The T-Test table displays the Input parameters such as the name of the active data set, filters, weight, split files, and the number of rows in the working data file. Missing Value Handling, Syntax, and Resources are described further.

Figure 2 shows the comparison of the mean accuracy value with algorithms Novel Support Vector Machine and Logistic Regression where the mean accuracy of Novel Support Vector Machine is about 69.5% and the mean accuracy value of the Logistic Regression is about 66.5%.

Discussion

Based on the results obtained by independent sample T-test analysis, the significance value is determined as 1.0, that means $p>0.05$ and shows insignificant due to the inconsistent dataset. The accuracy of SVM is 69.5% which is higher than LR accuracy 66.5%. This shows that SVM has betterment compared to LR for the loan approval process.

The risk variables have risen in tandem with the fast expansion of the economy and commercial transaction volumes. When such behavior goes unreported, it might lead to an increase in bad debt. We evaluate and contrast the measures of accuracy of the Novel Support Vector Machine with the classic Logistic Regression technique to solve all of the drawbacks of the current algorithms. Factors affecting the research (Benmouiza 2022) work are the predictive models that specify the comparison of two models with the best performance and accuracy. The analysis of both algorithms has been done with Table 1 representing the group statistics (Deng, Xing, and Cai 2022) and Table 4 representing the independent variables, and a bar graph which represents the comparison of the two algorithms with the accuracy percentages of 69.5% for Novel Support Vector Machine and Logistic Regression (Nallathambi et al. 2022) with an accuracy of 66.5%. Many studies are related to the Similar study of proposed research where (Sun, Yang, and Li 2022) the findings are, “Credit card fraud detection using AdaBoost and majority voting”, “Credit Card Fraud Detection: A Realistic Modeling and a Novel Learning Strategy", “Credit Card Fraud Detection: A Novel Approach Using Aggregation (Shinozaki et al. 2009) Strategy and Feedback Mechanism", “Prediction of Loan Status in Commercial Bank using Machine Learning Classifier".

Although the results of the study are better in both experimental and statistical analysis, there are certain limitations in the work. The evaluation of accuracy cannot provide a better outcome on larger data sets. However, the work can (Ross Quinlan 1993) be enhanced by applying optimization algorithm techniques, to achieve better accuracy. Feature selection algorithms can be used before classification to improve the accuracy of classifiers. The future scope of the study explains how it will be useful in the future for many applications with improved accuracy than other algorithms that don't take into account the necessary number of variables by carefully observing the credit risk while evaluating the credit score or to be precise, the approval score.

Conclusion

Background verification primarily falls into one of the major buckets of evaluation. The graph table displays the Input parameters such as the name of the active data set, filters, weight, split files, and the number of rows in the working data file. Missing Value Handling, Syntax, and Resources are described further. The mean accuracy rate of the Novel Support Vector Machine algorithm has been improved to 69.5% compared to Logistic Regression which is having around 66.5% mean accuracy. This suggests the proposed system provides an accurate analysis of creditworthiness for loan approval.

DECLARATIONS

Conflicts of interests

No conflicts of interests in the manuscript.

Authors Contribution

Author CHVS was involved in data collection, data analysis, and manuscript writing. Author SPC 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 that enabled us to complete the study.

1. Saveetha University
2. Saveetha Institute of Medical and Technical Sciences
3. Saveetha School of Engineering
4. Sree Vidya High School, Nandigama.

REFERENCES

1. Sukmandhani, Arief Agus, and IndrajaniSutedja. 2019. "Face Recognition Method for Online
2. Exams." 2019 International Conference on Information Management and Technology
3. (ICIMTech). https://doi.org/10.1109/icimtech.2019.8843831.
4. Ross Quinlan, J. 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann.
5. Benmouiza, Khalil. 2022. "Hourly Solar Irradiation Forecast Using Hybrid Local Gravitational Clustering and Group Method of Data Handling Methods." *Environmental Science and Pollution Research International*, April. https://doi.org/10.1007/s11356-022-20114-3.
6. Bolton, Christine. 2009. *Logistic Regression and Its Application in Credit Scoring*.
7. Deng, Yi, Chengyue Xing, and Ling Cai. 2022. "Building Image Feature Extraction Using Data Mining Technology." *Computational Intelligence and Neuroscience* 2022 (April): 8006437.
8. Devarajan, Yuvarajan, Beemkumar Nagappan, Gautam Choubey, Suresh Vellaiyan, and Kulmani Mehar. 2021. "Renewable Pathway and Twin Fueling Approach on Ignition Analysis of a Dual-Fuelled Compression Ignition Engine." *Energy & Fuels: An American Chemical Society Journal* 35 (12): 9930–36.
9. Dhanraj, Ganapathy, and Shanmugam Rajeshkumar. 2021. "Anticariogenic Effect of Selenium Nanoparticles Synthesized Using Brassica Oleracea." *Journal of Nanomaterials* 2021 (July). https://doi.org/10.1155/2021/8115585.
10. Kamath, S. Manjunath, K. Sridhar, D. Jaison, V. Gopinath, B. K. Mohamed Ibrahim, Nilkantha Gupta, A. Sundaram, P. Sivaperumal, S. Padmapriya, and S. Shantanu Patil. 2020. "Fabrication of Tri-Layered Electrospun Polycaprolactone Mats with Improved Sustained Drug Release Profile." *Scientific Reports* 10 (1): 18179.
11. Lee, Seong-Whan, and Alessandro Verri. 2003. *Pattern Recognition with Support Vector Machines: First International Workshop, SVM 2002, Niagara Falls, Canada, August 10, 2002. Proceedings*. Springer.
12. Muflikhah, Lailil, Nurul Hidayat, and Dimas Joko Hariyanto. 2019. "Prediction of Hypertention Drug Therapy Response Using K-NN Imputation and SVM Algorithm." *Indonesian Journal of Electrical Engineering and Computer Science*. https://doi.org/10.11591/ijeecs.v15.i1.pp460-467.
13. Nallathambi, Indumathi, Ramalakshmi Ramar, Denis A. Pustokhin, Irina V. Pustokhina, Dilip Kumar Sharma, and Sudhakar Sengan. 2022. "Prediction of Influencing Atmospheric Conditions for Explosion Avoidance in Fireworks Manufacturing Industry-A Network Approach." *Environmental Pollution* 304 (July): 119182.
14. Nandhini, Joseph T., Devaraj Ezhilarasan, and Shanmugam Rajeshkumar. 2020. "An Ecofriendly Synthesized Gold Nanoparticles Induces Cytotoxicity via Apoptosis in HepG2 Cells." *Environmental Toxicology*, August. https://doi.org/10.1002/tox.23007.
15. Parakh, Mayank K., ShriraamUlaganambi, Nisha Ashifa, Reshma Premkumar, and Amit L. Jain. 2020. "Oral Potentially Malignant Disorders: Clinical Diagnosis and Current Screening Aids: A Narrative Review." *European Journal of Cancer Prevention: The Official Journal of the European Cancer Prevention Organisation* 29 (1): 65–72.
16. Perumal, Karthikeyan, Joseph Antony, and Subagunasekar Muthuramalingam. 2021. "Heavy Metal Pollutants and Their Spatial Distribution in Surface Sediments from Thondi Coast, Palk Bay, South India." *Environmental Sciences Europe* 33 (1). https://doi.org/10.1186/s12302-021-00501-2.
17. Pham, Quoc Hoa, SupatChupradit, Gunawan Widjaja, Muataz S. Alhassan, Rustem Magizov, Yasser Fakri Mustafa, Aravindhan Surendar, AmirzhanKassenov, Zeinab Arzehgar, and WanichSuksatan. 2021. "The Effects of Ni or Nb Additions on the Relaxation Behavior of Zr55Cu35Al10 Metallic Glass." *Materials Today Communications* 29 (December): 102909.
18. Ross Quinlan, J. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann.
19. 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.
20. Shinozaki, Tsuguhiro, Mahito Morita, Hiroyuki Ohishi, and Kohei Furukawa. 2009. "A STUDY ON THE METHOD OF MAKING OCCURRENCE AND NONOCCURRENCE RULES OF DEBRIS FLOW THAT USES SVM AND ROUGH SET." *Doboku Gakkai Ronbunshuu F*. https://doi.org/10.2208/jscejf.65.448.
21. Siddig, Osama, Ahmed Farid Ibrahim, and SalaheldinElkatatny. 2021. "Application of Various Machine Learning Techniques in Predicting Total Organic Carbon from Well Logs." *Computational Intelligence and Neuroscience* 2021 (August): 7390055.
22. Sun, Zhenlong, Jing Yang, and Xiaoye Li. 2022. "Differentially Private Singular Value Decomposition for Training Support Vector

Machines." *Computational Intelligence and Neuroscience* 2022 (March): 2935975.
23. Tesfaye Jule, Leta, Krishnaraj Ramaswamy, Nagaraj Nagaprasad, Vigneshwaran Shanmugam, and Venkataraman Vignesh. 2021. "Design and Analysis of Serial Drilled Hole in Composite Material." *Materials Today: Proceedings* 45 (January): 5759–63.
24. Uganya, G., Radhika, and N. Vijayaraj. 2021. "A Survey on Internet of Things: Applications, Recent Issues, Attacks, and Security Mechanisms." *Journal of Circuits Systems and Computers* 30 (05): 2130006.
25. Zhu, Siyao, Cassandra Mitsinikos, Lisa Poirier, Takeru Igusa, and Joel Gittelsohn. 2021. "Development of a System Dynamics Model to Guide Retail Food Store Policies in Baltimore City." *Nutrients* 13 (9). https://doi.org/10.3390/nu13093055.

TABLES AND FIGURES

Table 1. Pseudocode for supporting SVM over logistic regression algorithm Based on the initialize, compute, update, sum of vectors compares the values.

Input: Bankruptcies $x_{i,}$ labels y_i
Output: Sum of vectors ,a array, b and SVM
Procedure: 1: Initialize:$a_{i,}$=0,f_i = -y 2:Compute:svc_loan = SVC().fit(xtrain_loan, ytrain) 3:Update:Loan status 4:Compute:svc_credit = SVC().fit(xtrain_credit, ytrain) 5: Until :vc_income = SVC().fit(xtrain_income, ytrain) 6:Update the threshold b 7:Store the status value 8:Update the data entry 9: Determine the datasets of credentials.

Table 2. PseudoCode for Logistic Regression which is an optimization algorithm for finding the local minimum of a differentiable function. Gradient descent is used to find values of a function parameter.

Input: v: Bank credentials T: Unique terms in all documents
Output: Accuracy
Procedure: for temp = list(find dataframe.columns) for dataframe = dataframe.dropna() wij= accuracy of approval Often $t_i$in document d_j End for document End for of term

Table 3. Comparison between significance level for SVM over Logistic Regression with the significance value has the standard deviation as 3.02.

	Algorithm	N	Mean	std.dev	Std.error mean
Accuracy	SVM	20	69.5000	3.02765	0.95743
	LR	20	66.5000	3.02765	0.95743

Table 4. Independent sample test for significance and standard error determination. The significance value obtained has an insignificant difference of 1.0 between the two groups for the selected dataset. The p-value = 1.0, Mean Difference = 3.00 and Confidence interval = (0.15 - 5.84).

		Levene's test for equality variance		**T-test for Equality of Means**		**T-test for Equality of Means**				
						Sig (2-tailed)	**Mean Difference**	**std**	**95.5% confidence interval of the**	
		F	**sig**	**t**	**df**				**Lower**	**Upper**
Accuracy	Equal variance assumed	0.0	1.0	2.21	18	0.04	3.00000	1.35401	.15534	5.84466
	Equal variance not assumed			2.21	18.0	0.04	3.00000	1.35401	.15534	5.84466

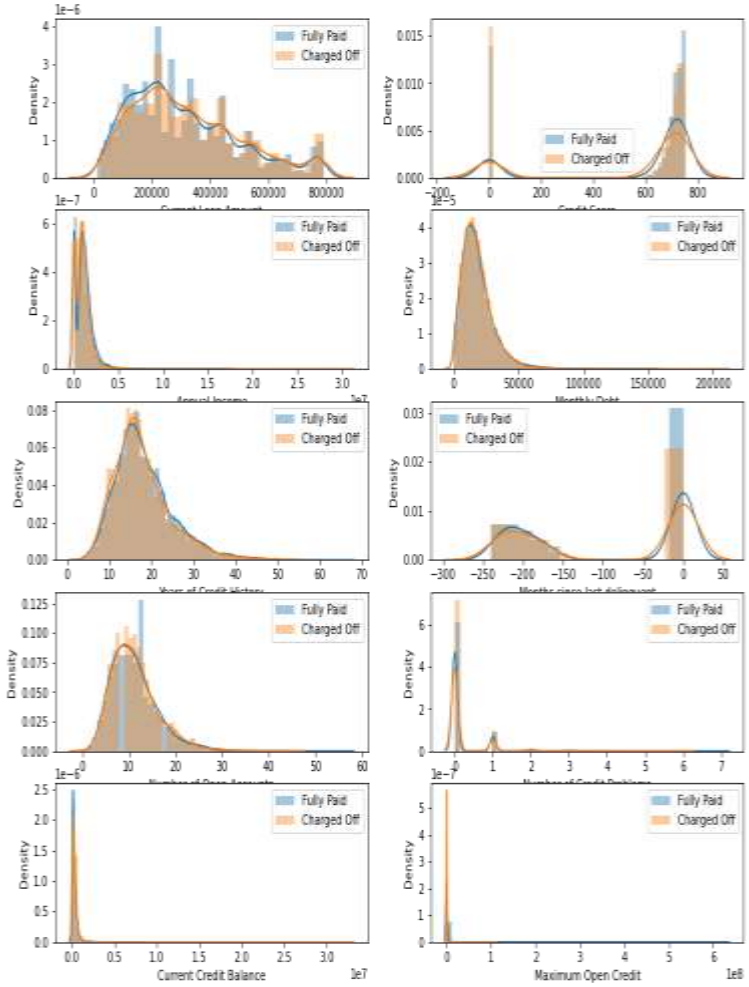

Fig. 1. The above graphs represent the variation of different people calculated with their credit score and we can also observe that the graph shows the current credit balance of the account.

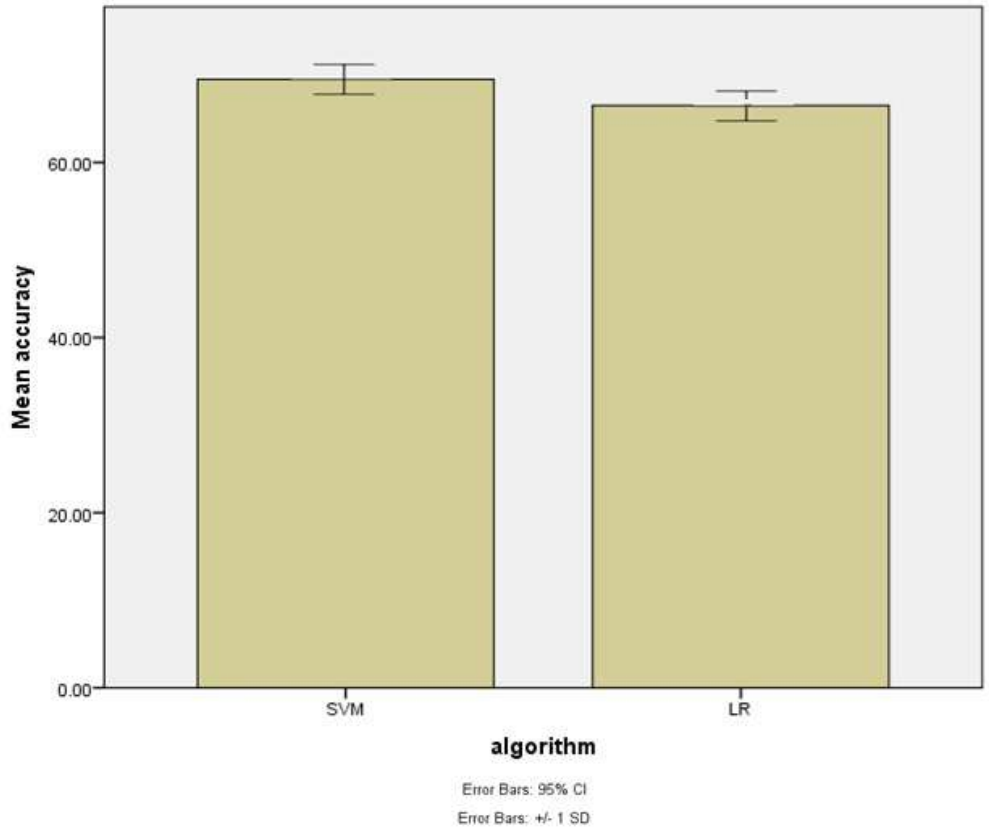

Fig. 2. The graph shows the comparison of the mean accuracy value with algorithms Novel Support Vector Machine (SVM) and Logistic Regression (LR), where the mean accuracy of Novel Support Vector Machine is about 69.5% and the mean accuracy value of the Logistic Regression is about 66.5%. X axis: SVM vs Logistic Regression Algorithm, Y axis: Mean accuracy of detection ± 1 SD.