Original Article

Efficient Prediction of Heart Disease using SVM Classification Algorithm and Compare its Performance with Linear Regression in Terms of Accuracy

B. Manoj Kumar[1], P S.Uma Priyadarsini[2]

[1]Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamilnadu, India, pin: 602105
[2*]Project Guider, Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamilnadu, India, pin: 602105

Abstract

Aim: The main objective of this research article is to employ the detection of heart disease by using Support vector machine (SVM) classifier in comparison with Linear regression (LR) model. **Materials & Methods:** The dataset used in this paper was collected from the UCI machine learning repository database. The sample size for the detection of heart disease was sample 60 (Group 1=30 and Group 2 =30) and calculation was performed utilizing G-power 0.8 with alpha and beta qualities of 0.05, 0.2 with a confidence interval of 95%. The detection of heart disease is performed by the Support Vector Machine (SVM) classifier with a number of samples (N=30) and Linear regression (LR) model with a number of samples (N=30). **Results:** The Support vector machine (SVM) classifier has a 90.43 percent higher accuracy rate when compared to the accuracy rate of the Linear regression (LR) model is 78.56 percent. The study has a significance value of p=0.021. **Conclusion:** Support vector machine (SVM) classifier provides better outcomes in accuracy rate when compared to Linear regression (LR) model for detection of heart disease.

Keywords: Image processing, Novel Support Vector Machine (SVM) classifier, Linear Regression (LR) model, accuracy rate, Heart Disease, Segmentation.

DOI:10.47750/pnr.2022.13.S04.171

INTRODUCTION

The human heart is the body's most crucial and important organ. Because even a slight miscalculation might result in weariness or death, diagnosing and forecasting heart problems requires higher precision, fineness, and accuracy(Singh and Kumar 2020). There are numerous heart-related deaths, and the number is steadily increasing(Gaziano et al. 2010). A classification method for heart disease surveillance is needed to solve the problem. This research uses the UCI machine learning repository dataset for training and testing to calculate the accuracy of novel support vector machine (SVM) and linear regression (LR) machine learning algorithms for predicting heart disease(Anggoro and Kurnia 2020). The proposed approach is estimated using parameters such as accuracy, specificity, and sensitivity.

Recently, numerous studies have been conducted in healthcare institutions employing various data mining approaches and machine learning techniques to develop disease detection systems(Palaniappan and Awang 2008; Khanna et al. 2015; Bhatia, Prakash, and Pillai 2008). IEEE Xplore distributed 67 examination papers, and Google Scholar tracked down 97 articles. The researchers in(Kumari and Godara n.d.) employed different classifiers such Decision tree, ANN, and Support Vector Machine and their testing findings revealed that Support Vector Machine had the best prediction

accuracy. K. Polaraju et al.(Polaraju, Prasad, and Others 2017) established a Multiple Regression Model for Identifying Heart Disease, which indicates that Multiple Linear Regression is suitable for forecasting heart disease risk. Marjia et al.(Sultana and Haider 2017) used KStar, j48, SMO, and Bayes Net, as well as WEKA software, to produce heart disease prediction utilizing KStar, j48, SMO, and Multilayer perceptron. Prediction and analysis of the incidence of Heart disease using Data Mining Techniques was suggested by Chala Beyene et al,(Beyene and Kamat 2018). For cardiac disease prediction, R. Sharmila et al.(Sharmila and Chellammal 2018) recommended using a non-linear classification system. For heart disease diagnosis using an optimal attribute set, it is proposed to employ big data techniques such as Hadoop Distributed File System (HDFS), Mapreduce, and SVM. Jayami Patel et al.(Patel 2015) proposed utilizing data mining and machine learning algorithms to detect heart disease. According to Ashwini Shetty et al.(Shetty and Naik 2016), a classification method for diagnosing heart illness using a patient's medical data should be established. Using the UCI machine learning dataset, which has 303 samples with 14 input features; Kumar et al.(Kumar, Koushik, and Deepak 2018) studied numerous machine learning and data mining methods and showed that SVM is the greatest among them.

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 method's fundamental flaw is that it is ineffective at detecting heart disease, and the algorithm is unable to determine the valuable and precise boundary of the heart disease zone. To overcome this limitation, this research proposed a novel Support Vector Machine that can predict heart illness based on variables such as sex, age, pulse rate, and others. When compared to the linear regression technique, the novel support vector machine algorithm that is utilized in this research study will provide the most accurate and dependable results.

MATERIALS AND METHODS

Dataset Description

This work was carried out in the Image Processing Laboratory, Department of Computer Science and Engineering, Saveetha School of Engineering. In this study, the dataset was collected from the UCI machine learning repository. 60 sample images were taken. The database is divided by the amount of 75% training and 25% testing. Two sets are taken and 30 data samples for each set, total number of samples considered are 60. Group 1 was a Linear regression (LR) algorithm and Group 2 was a novel Support vector machine (SVM) algorithm. The output is obtained by using Matlab software for the prediction of heart disease. The calculation is performed utilizing G-power 0.8 with alpha and beta qualities 0.05, 0.2 with a confidence interval at 95%(Chen et al. 2007).

Linear Regression

The linear regression (LR) algorithm is a supervised learning technique that predicts the outcome using known parameters that are connected with the output. The relationship between the independent and dependent variables is the basis for it. The variables "x" and "y" are independent and dependent variables, respectively, and the relationship between them is represented by a line equation, which is linear in nature, hence the name "linear regression."
The parameters of X0 and X1 should be chosen to provide the least amount of inaccuracy. There are various types of error measurements that can be used to evaluate the model. If the quantity of squared error is used as a measurement to evaluate the model, the following equation is used to calculate the error.

Pseudocode

Input: Heart disease_Input Features
Assign Training and Testing Dataset of Heart disease Dataset
Output: Prediction of Heart disease
Step 1: Require
Step 2: Determine the Initial Vector
Step 3: Apply zero to the first and second moment vectors
Step 4: Set the t value of the timestamp to zero.
Step 5: Perform while (conditioning)
Step 6: At timestamp t, obtain gradients with respect to stochastic unbiased.

Step 7: Renovate the first and second moment estimates that are skewed.
Step 8: For the first and second moment estimates, calculate bias-connected.
Step 9: Retain the results of the classification.

Support Vector Machine (SVM)

Gender categorization problems have been successfully solved using the support vector machine (SVM). In an SVM classifier, the separation hyper plane is chosen to minimize the anticipated generalization error of the unobservable sample data. SVM is a sophisticated classifier that can tell the difference between two groups of people. The test image is assigned to the group with the largest separation to the training's closest point by SVM. The SVM training procedure develops a strategy that can identify whether an input image belongs to this class or not. To find an accurate decision border, SVM requires a considerable amount of training data, which increases the computational cost. The SVM is a learning algorithm for categorisation. It aims to find the optimum separation hyperplane for unobserved sequences with the lowest possible predicted classification error. For linearly non-separable data, the input is transported to a high-dimensional feature set where they can be distinguished by a hyperplane. To achieve this projection onto a high-dimensional feature space, kernels are required.

Pseudocode

Input: heart disease_Input Features
Assign training and testing dataset for heart disease
Output: Classification of heart disease as a result of the output
Function: Support_Vector_Machine(Input features F, Label vector V=[1.....n])
Step 1: Decide on the optimal cost and gamma value.
Step 2: Perform while (conditioning)
Step 3: For each set number of input file features, run the training step.
Step 4: Run the classification step for a set number of features in the input file.
Step 5: Come to an end whilst
Step 6: Submit the heart disease classification results.

Statistical Analysis

Matlab software is used to generate the results(Knight 2019). A monitor with a resolution of 1024x768 pixels was required to train these datasets (10th gen, i5, 12GB RAM, 500 GB HDD). The software programme IBM SPSS is employed in this study for statistical analysis(Frey 2017). The independent sample t test was used to determine the mean, standard deviation, and standard error mean statistical significance between the groups, and then the two groups were compared using SPSS software to obtain accurate values for the two different s, which were then used with the graph to calculate the significant value with maximum accuracy (90.43 percent), mean value (97 percent), and standard deviation value (1.37848). Accuracy is a dependent variable, while SVM and LR are independent variables.

RESULTS

The accuracy rate of the SVM classifier is compared to the LR classifier in Figure 1. The SVM classifier has a higher accuracy rate of 90.43 when compared to the LR classifier, which have 78.56 respectively. The SVM classifier is significantly different from the LR classifier ($p<0.05$ independent sample test). SVM, LR accuracy rates are plotted on the X-axis. Y-axis: Mean accuracy rate for keyword identification, ± 1 SD with 95 percent confidence interval.
Table 1 presents the evaluation metrics of the comparison of the SVM classifier with the LR classifier. The SVM classifier has a 90.43 accuracy rate, whereas the LR classifier has 78.56, respectively. In all parameters, the SVM classifier outperforms the LR in the classification of heart disease, with a higher accuracy rate.

Table 2 displays the statistical computations for SVM, LR classifier, such as mean, standard deviation, and standard error mean. In the t-test, the accuracy rate parameter is used. The SVM classifier has a mean accuracy rate of 90.43, while the LR classifier has 78.56, respectively. The SVM classifier has a standard deviation of 1.37848, while LR has a standard deviation of 2.38498 respectively. The SVM classifier has a Standard Error Mean of 0.67283, while LR has a Standard Error Mean of 1.73823 respectively.

Table 3 shows the statistical computations for independent samples of SVM compared to the Linear regression classifier. The accuracy rate has a significance level of 0.021. The SVM classifier is compared to LR using an Independent samples T-test with a confidence interval of 95 percent and a threshold of significance of 0.37283. The significance level is 0.001, the significance level is two-tailed, the mean difference, the standard error difference, and the lower and upper interval difference are all included in this independent sample test.

DISCUSSION

For evaluating heart disease classification, this study offered two different algorithms and prediction models. According to the results of the studies, Support vector machine (90.43 percent) outperforms Decision Tree (70.42 percent), Naive Bayesian algorithm (74.15 percent), and Random forest (74.15 percent) in terms of accuracy (65.95 percent).

For the past five years, there have been numerous studies published in the literature. By integrating the properties of the Linear Method (LM) and Random Forest, the authors presented the hybrid HRFLM technique (RF). They were able to forecast with an accuracy of 88.4%(Mohan, Thirumalai, and Srivastava 2019). In 2020, the authors used the Starlog and Cleveland heart disease datasets to test seven different intelligent techniques for predicting coronary heart disease, and in their comparison analysis, the deep neural network outperformed and achieved an accuracy of 98.15 percent with the Starlog dataset, while SVM achieved an accuracy of 97.36 percent with the Cleveland dataset(Ayon, Islam, and Hossain 2020). S. Seema et al.(Deepika and Seema 2016) focus on strategies that use Naive Bayes, D-Tree, SVM, and ANN to predict chronic disease by mining data from past health records. A comparison analysis of classifiers is conducted in order to determine which performs better at a more accurate rate. SVM had the best performance in this experiment. Purushottam et al.(Purushottam, Saxena, and Sharma 2016) developed a data-mining-based approach for predicting cardiac disease. This system aids medical practitioners in making informed decisions based on a set of criteria. It provides 86.3 percent accuracy in the testing phase and 87.3 percent accuracy in the training phase by testing and training a specific parameter.

The disadvantage of the proposed study is its limited sample size, which makes statistical significance for any of the goals extremely challenging. The performance of the support vector machine can be improved in the future by employing a combination of different approaches and data trimming. Other machine learning approaches can be used to improve accuracy.

CONCLUSION

The proposed model exhibits the Support vector machine (SVM) classifier and Linear regression (LR) model, in which the Support vector machine (SVM) classifier has the highest accuracy values. The accuracy rate of the Support vector machine (SVM) classifier is 90.43% higher compared with the Linear regression (LR) model that has an accuracy rate of 78.56% in detection of heart disease with improved accuracy rate.

DECLARATION

Conflicts of Interest

No conflict of interest in this manuscript

Authors Contributions

BMK was involved in data collection, data analysis & manuscript writing. Author PSU was involved in conceptualization, data validation, and critical review of manuscripts.

Acknowledgment

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 successfully carrying out this work.

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

REFERENCES

1. Anggoro, Dimas Aryo, and Naqshauliza Devi Kurnia. 2020. "Comparison of Accuracy Level of Support Vector Machine (SVM) and K-Nearest Neighbors (KNN) Algorithms in Predicting Heart Disease." *Aquatic Microbial Ecology: International Journal* 8 (5). https://www.academia.edu/download/63554423/ijeter3285202020200607-113398-1w5oaj3.pdf.
2. Ayon, Safial Islam, Md Milon Islam, and Md Rahat Hossain. 2020. "Coronary Artery Heart Disease Prediction: A Comparative Study of Computational Intelligence Techniques." *IETE Journal of Research*, January, 1–20.
3. Beyene, Chala, and Pooja Kamat. 2018. "Survey on Prediction and Analysis the Occurrence of Heart Disease Using Data Mining Techniques." *International Journal of Pure and Applied Mathematics: IJPAM* 118 (8): 165–74.
4. Bhatia, Sumit, Praveen Prakash, and G. N. Pillai. 2008. "SVM Based Decision Support System for Heart Disease Classification with Integer-Coded Genetic Algorithm to Select Critical Features." In *Proceedings of the World Congress on Engineering and Computer Science*, 34–38. iaeng.org.
5. Chen, J., Y. Xing, G. Xi, J. Chen, J. Yi, and D. Zhao. 2007. "A Comparison of Four Data Mining Models: Bayes, Neural Network, SVM and Decision Trees in Identifying Syndromes in Coronary Heart Disease." *Symposium on Neural …*. https://link.springer.com/chapter/10.1007/978-3-540-72383-7_148.
6. Deepika, Kumari, and S. Seema. 2016. "Predictive Analytics to Prevent and Control Chronic Diseases." In *2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT)*, 381–86. ieeexplore.ieee.org.
7. 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.
8. 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.
9. Frey, Felix. 2017. "SPSS (Software)." *The International Encyclopedia of Communication Research Methods*, November, 1–2.
10. Gaziano, Thomas A., Asaf Bitton, Shuchi Anand, Shafika Abrahams-Gessel, and Adrianna Murphy. 2010. "Growing Epidemic of Coronary Heart Disease in Low- and Middle-Income Countries." *Current Problems in Cardiology* 35 (2): 72–115.
11. 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.
12. Khanna, Divyansh, Rohan Sahu, Veeky Baths, and Bharat Deshpande. 2015. "Comparative Study of Classification Techniques (SVM, Logistic Regression and Neural Networks) to Predict the Prevalence of Heart Disease." *International Journal of Machine Learning and Computing* 5 (5): 414.
13. Knight, Andrew. 2019. *Basics of MatLab® and beyond*. Chapman and Hall/CRC.
14. Kumari, Milan, and Sunila Godara. n.d. "Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction 1." Accessed June 3, 2022. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.219.6038.
15. Kumar, M. Nikhil, K. V. S. Koushik, and K. Deepak. 2018. "Prediction of Heart Diseases Using Data Mining and Machine Learning Algorithms and Tools." *International Journal of Scientific Research in Computer Science, Engineering and Information Technology* 3 (3): 887–98.
16. Mohan, Senthilkumar, Chandrasegar Thirumalai, and Gautam Srivastava. 2019. "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques." *IEEE Access* 7: 81542–54.
17. 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.
18. Palaniappan, Sellappan, and Rafiah Awang. 2008. "Intelligent Heart Disease Prediction System Using Data Mining Techniques." In *2008 IEEE/ACS International Conference on Computer Systems and Applications*, 108–15.
19. Parakh, Mayank K., Shriraam Ulaganambi, 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.
20. Patel, Jaymin. 2015. "Prof. Tejal Upadhyay, Dr. Samir Patel,"." *Heart Disease Prediction Using Machine Learning and Data Mining Technique* 7 (1Sept): 2016.
21. 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.
22. Pham, Quoc Hoa, Supat Chupradit, Gunawan Widjaja, Muataz S. Alhassan, Rustem Magizov, Yasser Fakri Mustafa, Aravindhan Surendar, Amirzhan Kassenov, Zeinab Arzehgar, and Wanich Suksatan. 2021. "The Effects of Ni or Nb Additions on the Relaxation Behavior of Zr55Cu35Al10 Metallic Glass." *Materials Today Communications* 29 (December): 102909.
23. Polaraju, K., D. Durga Prasad, and Others. 2017. "Prediction of Heart Disease Using Multiple Linear Regression Model." *International Journal of Engineering Development and Research* 5 (4): 2321–9939.
24. Purushottam, Kanak Saxena, and Richa Sharma. 2016. "Efficient Heart Disease Prediction System." *Procedia Computer Science* 85 (January): 962–69.
25. 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.
26. Sharmila, R., and S. Chellammal. 2018. "A Conceptual Method to Enhance the Prediction of Heart Diseases Using the Data Techniques." *International Journal of Computer Science and Engineering* 6 (4): 21–25.

27. Shetty, Ashwini, and Chandra Naik. 2016. "Different Data Mining Approaches for Predicting Heart Disease." *Int J Innov Res Sci Eng Technol* 5 (9): 277–81.
28. Singh, Archana, and Rakesh Kumar. 2020. "Heart Disease Prediction Using Machine Learning Algorithms." In *2020 International Conference on Electrical and Electronics Engineering (ICE3)*, 452–57. ieeexplore.ieee.org.
29. Sultana, Marjia, and Afrin Haider. 2017. "Heart Disease Prediction Using WEKA Tool and 10-Fold Cross-Validation." In *The Institute of Electrical and Electronics Engineers*, 17–33.
30. 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.
31. 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.

TABLES AND FIGURES

Table 1. The evaluation metrics of the SVM classifier with the LR classifier has been calculated. The SVM classifier has a 90.43 accuracy rate, whereas the LR classifier has 78.56, respectively. In all parameters, the SVM classifier outperforms the LR in the classification of heart disease, with a higher accuracy rate.

ITERATION NO.	Support Vector Machine Classification(%)	Linear Regression(%)
1	92.3	88.6
2	91.9	87.2
3	89.5	86.3
4	90.4	84.5
5	91.6	86.5
6	92.4	85.4
7	91.4	88.6
8	89.3	86.5
9	91.2	87.4
10	90.3	86.8

Table 2. The statistical calculation such as Mean, standard deviation and standard error Mean for SVM classifier and Linear regression (LR) model. The accuracy rate parameter used in the t-test. The mean accuracy rate of the SVM classifier is 90.43 and LR model is 78.56. The Standard Deviation of SVM classifier is 1.37848 and the LR model is 2.38498. The Standard Error Mean of the SVM classifier is 0.67283 and the LR model is 1.73823.

GROUP	N	Mean(%)	Std.Deviation	Std.Error Mean
Novel Support Vector Machine Classification	10	91.030	1.1056	.3496
Linear Regression	10	86.780	1.2770	.4038

Table 3. The statistical calculations for independent samples test between SVM classifier and Linear regression (LR) model. The sig. for accuracy rate is 0.021. Independent samples T-test is applied for comparison of SVM classifier and Linear regression (LR) model with the confidence interval as 95% and level of significance as 0.37283. This independent sample test consists of significance as 0.001, significance (2-tailed), Mean difference, standard error difference, and lower and upper interval difference.

		Levene's Test for Equality o Variance		T-test for Equality of Means						
									95% Confidence Interval of the Difference	
	Equal Variance	F	Sig	t	df	Sig(2-tailed)	Mean Difference	Std.Er ror Difference	Lower	Upper
Efficiency	**Assumed**	.003	<.001	7.957	18	.001	4.2500	.5341	3.1278	5.3722
	Not Assumed			7.957	17.639	.002	4.2500	.5341	3.1362	5.3738

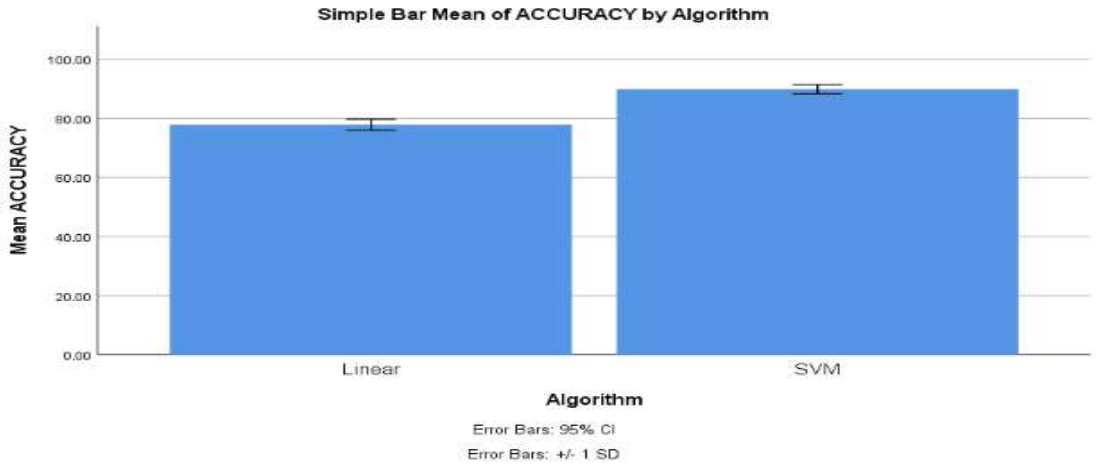

Fig. 1. Simple Bar graph for SVM classifier accuracy rate is compared with Linear regression (LR) model. The SVM classifier is higher in terms of accuracy rate 90.43 when compared with Linear regression (LR) model 78.56. Variable results with its standard deviation ranging from 100 lower to 150 higher in SVM classifier where Linear regression (LR) model standard deviation ranging from 200 lower to 300 higher. There is a significant difference between SVM classifier and Linear regression (LR) model ($p<0.05$ Independent sample test). X-axis: Linear regression (LR) model accuracy rate vs SVM classifier Y-axis: Mean of accuracy rate, for identification of keywords ± 1 SD with 95 % CI.