Support Vector Machine Algorithm with Improved Precision vs Logistic Regression Algorithm for Prediction of Cardiovascular Disease

Rajvardhan Gadde[1], Neelam Sanjeev Kumar[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: To compare and achieve the optimal way for the prediction of Novel cardiovascular condition accurately, with fewer errors between Logistic Regression and Support Vector Machine classifiers. **Materials and Methods:** Data collection containing various data points for predicting Novel cardiovascular disease from UCI machine learning repository. Classification is performed by Logistic Regression classifier (N=20) over Support Vector Machine (N=20) total sample size calculation is done through clinical.com. The accuracy was calculated using Matlab software and the outputs are graphed using SPSS software. **Results:** comparison of accuracy rate is done by independent sample test using SPSS software. There is a statistical indifference between Logistic Regression (LR) and Support Vector Machine(SVM). Support Vector Machine algorithm (87.38%) showed better results in comparison to Logistic Regression (74.73%). **Conclusion:** Support Vector Machine algorithm appears to give better accuracy than Logistic Regression algorithm for the prediction of Novel Cardiovascular Disease.

Keywords: Novel Cardiovascular Disease, Machine Learning, Logistic Regression Algorithm, Support Vector Machine Algorithm,Accuracy,Precision

INTRODUCTION

Heart disease is one of the severe health disorders which results in adverse health conditions if it is left untreated. In this era where the cardiac disease is peaking each year and causing many health issues and which in return reduces the productivity of every individual human. The diagnosis of Novel cardiovascular disease using the traditional method is a very difficult technique and with the rise of the number of Novel cardiovascular diseases, the testing is also being difficult and making cardiovascular one of the essential diagnostic areas where machine learning (ML) has to be introduced (Li et al. 2020). Many factors are responsible for Novel cardiovascular diseases for example personal and professional habits and genetic occurrences. By improving the prediction techniques we can improve the prediction accuracy which saves many lives (Bagheri et al. 2021). Data collection and testing and training those samples are some of the best methods for improving the accuracy of the prediction of Novel cardiovascular diseases (Shah, Patel, and Bharti 2020).

The primary point of this research finding is to get the optimal algorithm for the prediction of Novel cardiovascular disease by the comparison between logistic regression and Support VectorMachine classifiers (Marbaniang, Choudhury, and Moulik 2020). The prediction accuracy of Novel cardiovascular diseases can be increased by the improvement of the machine learning techniques and the usage of the previous data collected (Kumar, Gyawali, and Agarwal 2020)(A et al. 2020).

About 150 Science direct and 47 IEEE Explorer articles were found similar to this work in the last 5 years and have a clear report of developed algorithms and models using ML algorithms such as SVM,Naive Bayes, LR, Neural Network, Random Forest algorithms to predict and evaluate the accomplishment of every algorithm in terms of sensitivity, precision and accuracy in the prediction of Novel cardiovascular diseases (Arunachalam 2020; Subha 2016). In this paper, the major aim is to evaluate the validity of every algorithm in terms of accuracy, sensitivity, precision, and specificity and to perceive the best accuracy obtaining algorithm for the prediction of Novel cardiovascular disease(Subha 2016). Research work proposed a machine learning algorithm comparison of various classifiers to predict and reduce deaths due to Novel cardiovascular diseases (A et al. 2020). Accuracy comparison is done over different classifiers Naive Bayes, LR, Neural Network, Random Forest, diagnoses, and SVM on UCI Machine Learning Repository data set. All these classifiers are executed in simulated environments using Matlab data mining tools [Citation error]. The executed results depict high accuracy by the SVM with an accuracy of 87.38% and with the least error rate whereas the Logistic Regression algorithm got 74.73%. the precision values of the SVM and LR also are 90.85% and 76.72% respectively, followed by recall 84.52% by SVM and 74.74% by Logistic Regression and F1 values 87.55% and 75.63% are also out ruled by the Support Vector Machine classifier with higher values than the Logistic.

The main reason for the fluctuations and the variations in the accuracies is the data redundancies and data overlapping. The preprocessing of the data is very important for obtaining a high accuracy prediction. Data mining has a vital part in increased accuracy. When the Sample Size of the data training is increased the machine learning capability can be raised in return the accuracy of the prediction of Novel cardiovascular diseases increases. The authors are well versed in ML and data learning technologies. The principal point of the work is to look at Logistic regression and Support Vector Machine classifiers and finding the high accuracy yielding algorithm for the prediction of Novel cardiovascular diseases.

MATERIALS AND METHODS

The review was completed at the University recreation research center, Saveetha School of Engineering, SIMATS, Chennai. In the ongoing paper, the dataset was taken from UCI Machine Learning Repository Novel cardiovascular disease dataset. This data set consists of various features of the patients and different parameters of the patients in the given data set using the description of the various

features in the form of columnar attributes. There is visualization and analysis for support.

The data was donated by the UCI Machine Learning Repository and this includes all the parameters and the features which are required for the prediction, analysis, and evaluation of Novel cardiovascular diseases such as age and various heart parameters. This data is divided into two different groups. The sample size calculation was done using previous study results by clinical.com by keeping CI at 95%, alpha error-threshold by 0.05, enrollment ratio as 0:1, and power at 80%. Sample preparation is carried out for LR and SVM classifiers for the data collected from the UCI Repository .

The Logistic Regression algorithm is a probabilistic ML Algorithm that is empoy for decision-making tasks (Rani et al. 2018)[Citation error]. Logistic Regression will approximate the independence between the features of the dataset rules. But the Support Vector Machine algorithm is the higher accuracy giving algorithm which uses supervised learning and has excellent accuracy and classification performance. Support vector machine uses non -linear mapping to vary the training data to a greater dimension. The hyperplanes are selected by the Support Vector Machine algorithm (Jayadeva, Khemchandani, and Chandra 2016).

Group 1 is Logistic Regression and with N value 20 and group 2 is SVM with N value 20, the total sample size is 40. A sample dataset of both Logistic Regression and Support Vector Machine are exported to the Microsoft Excel Sheet for importing to the Matlab as input. Matlab 2021a software has to be installed on the PC for training the source dataset. Both Logistic Regression and SVM algorithms are employed to train the sample groups. A confusion matrix is obtained and True Positive(TP), False Positive(FP),True Negative(TN), and False Negative (FN) values are inscribed. Sensitivity (%),Accuracy (%) and precision (%) values are calculated from the resulting confusion matrix.

Statistical Analysis

The software used here for the statistical analysis is IBM SPSS V28.0.0.0 (190). Accuracy, precision, recall, and f1 Comparison of Logistic Regression with Support Vector Machine algorithm were done in this software. As the variables are independent of each other an independent sample T-test was fetched out to get mean values of accuracy, precision, recall, and F1 between two groups, and performance comparison between the two groups is performed.

RESULTS

In this research work of cardiovascular diseases prediction by Logistic Regression and Support Vector Machine on UCI Machine Learning Repository, the results depict to produce the same variable results with the accuracy of 74.73% and 87.38%, precision 76.72% and 90.85%, recall 74.74% and 84.52% and F1 75.63% and 87.55% respectively.

Table 2 shows the comparison of mean accuracy, mean sensitivity, and mean precision values of Logistic Regression and SVM. SVM shows higher values in terms of accuracy, sensitivity, and precision. Variable results with a

precision-90.85%,accuracy rate of 87.38%, recall value 84.52%, and F1 value of 87.55%. Whereas results of Logistic Regression are with an accuracy of 74.73%, precision of 76.72%, recall value of 74.74%, and F1 value of 75.63%.The Logistic Regression algorithm had less accuracy, precision, recall and F1 when compared to the Support Vector Machine algorithm as shown in Table 1a and Table 1b. The descriptive statistics of table 2 shows that the Support Vector Machine algorithm had less error when compared to the Logistic Regression algorithm.

Independent sample T-test results show that there is a statistically insignificant difference in accuracy ($P<0.001$), precision ($P<0.001$), recall ($P<0.001$), and F1($P<0.001$) as shown in table 3. Bar Chart constitutes the contrast of mean recall,mean precision and mean accuracy, mean F1 values of Logistic Regression and SVM as shown in Fig 1. Figure 2a and Fig. 2b represents the confusion matrix of LR and SVM respectively.

DISCUSSION

In this research paper for the prediction of cardiovascular diseases, we observed Support vector Machine had performed better with the precision of 90.85% and accuracy of 87.38%, recall 84.52%, and F1 value of 87.55% when compared to Logistic Regression are with precision of 76.72% and an accuracy of 74.73%, recall value of 74.74%, and F1 value of 75.63%. Although not statistically significant, the significant difference appears to have slightly increased table 3. Machine Learning has an important part in the early diagnosis of cardiovascular ailment. Moreover, preprocessing of the data will increase the prediction of cardiovascular diseases better.

Related works are done by many researchers [Citation error]proposed using similar comparison and by using machine learning algorithms and the main aim is to accurately evaluate the model in terms of precision,specificity, sensitivity, accuracy,and F- measure. Another study was done by [Citation error] this paper the author implemented ML calculation for the expectation of cardiovascular infections and by using a cardiovascular dataset that resulted from Accuracy, sensitivity, specificity, and MCC. A paper by [Citation error] used a similar feature section by using similar machine learning algorithms in which Logistic Regression had shown a lower accuracy value than the Support Vector Machine algorithm of 86.88% for the prediction of cardiovascular disease (Kumar, Gyawali, and Agarwal 2020). A comparative study of various classifiers was done in this paper (Jiang 2020) and the results reach the highest accuracy over the UCI Machine Learning Repository dataset.

The major factors that are affecting the accuracy are data redundancies and depending on the data size the accuracy may be varied. Further increase in the sample size will be yielding better F1 values, precision, recall, and accuracy. Preprocessing of the data is much needed for the optimal results for the prediction of cardiovascular diseases.

Limitation of this development of an efficient classification system that combines the effectiveness of the best

accuracy obtained for the improvement of the prediction. A large dataset of real-time applications paired with other ML algorithms and Deep Learning may improve the accuracy in future and the overall performance of the output. Overall, the findings of this study are highly promising for the future.

CONCLUSION

In this study of prediction of cardiovascular diseases, SVM has a higher accuracy of 87.38% than the Logistic Regression which has an accuracy of 74.73%. Support vector Machine had performed better with an accuracy of 87.38%, the precision of 90.85%, recall of 84.52%, and F1 value of 87.55% when compared to Logistic Regression are with an accuracy of 74.73%, precision of 76.72%, recall value of 74.74%, and F1 value of 75.63%. The performance of these algorithms can be increased with the increase of the data size.

DECLARATION

Conflicts of Interest

No conflict of interest in this manuscript

Author Contributions

Author GVSC was involved in data collection, data analysis & manuscript writing. The author NSK was involved in conceptualization, data validation, and critical review of manuscripts.

Acknowledgments

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.Venus Electronics Tamilnadu

2.Saveetha University

3.Saveetha Institute of Medical And Technical Sciences

4.Saveetha School of Engineering

REFERENCES

1. A, Jaya Lakshmi, Lakshmi A. Jaya, S. Venkatramaphanikumar, and Venkata Krishna Kishore Kolli. 2020. "Prediction of Cardiovascular Risk Using Extreme Learning Machine-Tree Classifier on Apache Spark Cluster." *Recent Advances in Computer Science and Communications*. https://doi.org/10.2174/2666255813999200904163404.
2. Arunachalam, Siddhika. 2020. "Cardiovascular Disease Prediction Model Using Machine Learning Algorithms." *International Journal for Research in Applied Science and Engineering Technology*. https://doi.org/10.22214/ijraset.2020.6164.
3. Bagheri, Ayoub, T. Katrien J. Groenhof, Folkert W. Asselbergs, Saskia Haitjema, Michiel L. Bots, Wouter B. Veldhuis, Pim A. de Jong, and Daniel L. Oberski. 2021. "Automatic Prediction of Recurrence of Major Cardiovascular Events: A Text Mining Study Using Chest X-Ray Reports." *Journal of Healthcare Engineering* 2021 (July): 6663884.
4. Jayadeva, Reshma Khemchandani, and Suresh Chandra. 2016. *Twin Support Vector Machines: Models, Extensions*

and Applications. Springer.
5. Jiang, Shu. 2020. *Heart Disease Prediction Using Machine Learning Algorithms*.
6. Kumar, Ashutosh, Rahul Gyawali, and Sonali Agarwal. 2020. "Cardiovascular Disease Prediction Using Machine Learning Tools." *Machine Intelligence and Signal Processing*. https://doi.org/10.1007/978-981-15-1366-4_35.
7. Li, Jian Ping, Amin Ul Haq, Salah Ud Din, Jalaluddin Khan, Asif Khan, and Abdus Saboor. 2020. "Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare." *IEEE Access*. https://doi.org/10.1109/access.2020.3001149.
8. Marbaniang, Ibashisha A., Nurul Amin Choudhury, and Soumen Moulik. 2020. "Cardiovascular Disease (CVD) Prediction Using Machine Learning Algorithms." *2020 IEEE 17th India Council International Conference (INDICON)*. https://doi.org/10.1109/indicon49873.2020.9342297.
9. Rani, K. Sandhya, K. Sandhya Rani, M. Sai Chaitanya, G. Sai Kiran, Dhanekula Institute of Engineering and Technology, Ganguru, Vijayawada, Andhra Pradesh, and India. 2018. "A Heart Disease Prediction Model Using Logistic Regression By Cleveland DataBase." *International Journal of Trend in Scientific Research and Development*. https://doi.org/10.31142/ijtsrd11402.
10. Shah, Devansh, Samir Patel, and Santosh Kumar Bharti. 2020. "Heart Disease Prediction Using Machine Learning Techniques." *SN Computer Science*. https://doi.org/10.1007/s42979-020-00365-y.

TABLES AND FIGURES

Table 1a. Cardiovascular Disease samples using Logistic Regression Algorithm

Sample	Accuracy	Precision	Recall	F1
1	0.8	0.782609	0.857143	0.818182
2	0.725	0.777778	0.666667	0.717949
3	0.775	0.8	0.761905	0.780488
4	0.714286	0.761905	0.695652	0.727273
5	0.75	0.789474	0.714286	0.75
6	0.725	0.75	0.714286	0.731707
7	0.725	0.777778	0.666667	0.717949
8	0.725	0.727273	0.761905	0.744186

9	0.75	0.761905	0.761905	0.761905
10	0.707317	0.75	0.681818	0.714286
11	0.75	0.761905	0.761905	0.761905
12	0.775	0.772727	0.809524	0.790698
13	0.75	0.789474	0.714286	0.75
14	0.8	0.782609	0.857143	0.818182
15	0.725	0.75	0.714286	0.731707
16	0.75	0.761905	0.761905	0.761905
17	0.75	0.761905	0.761905	0.761905
18	0.75	0.761905	0.761905	0.761905
19	0.75	0.761905	0.761905	0.761905
20	0.75	0.761905	0.761905	0.761905

Table 1b. Cardiovascular Disease samples using Support Vector Machine Algorithm

Sample	Accuracy	Precision	Recall	F1
1	0.85	0.894737	0.809524	0.85
2	0.9	0.947368	0.857143	0.9
3	0.825	0.85	0.809524	0.829268
4	0.9	0.947368	0.857143	0.9
5	0.85	0.894737	0.809524	0.85
6	0.9	0.947368	0.857143	0.9
7	0.85	0.894737	0.809524	0.85
8	0.9	0.947368	0.857143	0.9
9	0.875	0.9	0.857143	0.878049
10	0.85	0.857143	0.857143	0.857143
11	0.85	0.894737	0.809524	0.85

12	0.9	0.947368	0.857143	0.9
13	0.875	0.9	0.857143	0.878049
14	0.875	0.9	0.857143	0.878049
15	0.9	0.947368	0.857143	0.9
16	0.875	0.9	0.857143	0.878049
17	0.875	0.9	0.857143	0.878049
18	0.875	0.9	0.857143	0.878049
19	0.875	0.9	0.857143	0.878049
20	0.875	0.9	0.857143	0.878049

Table. 2 comparison of mean recall,mean precision, mean accuracy and mean F1 between Logistic Regression and Support Vector Machine.

Group Statistics					
	Group	**N**	**Mean**	**Std.Deviation**	**Std.Error Mean**
Accuracy	Logistic Regression	20	.7473	.02547	.00570
	Support Vector Machine	20	.8738	.02218	.00496
Precision	Logistic Regression	20	.7672	.01702	.00381
	Support Vector Machine	20	.9085	.02943	.00658
Recall	Logistic Regression	20	.7474	.05343	.01195
	Support Vector Machine	20	.8452	.02116	.00473
F1	Logistic Regression	20	.7563	.02959	.00662
	Support Vector Machine	20	.8755	.02143	.00479

Table. 3 Independent sample T-test in predicting the accuracy, precision, recall, and F1 of cardiovascular disease prediction using Logistic Regression and Support Vector Machine classifiers. There appears to be a statistically insignificant difference ($p<0.001$) in both the classifiers.

INDEPENDENT SAMPLE TEST									
		LEVENE'S TEST FOR EQUALITY OF VARIANCES		T-TEST FOR EQUALITY OF MEANS					
						SIGNIFICANCE		95% CONFIDENCE INTERVAL OF THE DIFFERENCE	
		F	SIG	T	DF	ONE-SIDED P	STD.ERROR DIFFERENCE	LOWER	UPPER
Accuracy	**Equal Variance Assumed**	.12	.729	-16.73	38	<.001	.00755	-.14171	-.11113
	Equal Variance is not Assumed			-16.73	37.29	<.001	.00755	-.14172	-.11112
Precision	**Equal Variance Assumed**	4.94	.032	-18.58	38	<.001	.00760	-.15666	-.12588
	Equal Variance is not Assumed			-18.58	30.42	<.001	.00760	-.15678	-.11575
Recall	**Equal Variance Assumed**	9.18	.004	-0.61	38	<.001	.01285	-.12381	-.07178
	Equal Variance is not Assumed			-7.61	24.81	<.001	.01285	-.12427	-.07132
F1	**Equal Variance Assumed**	1.14	.292	-14.59	38	<.001	.00817	-.13587	-.10271
	Equal Variance is not Assumed			-14.59	34.63	<.001	.00817	-.13583	-.10265

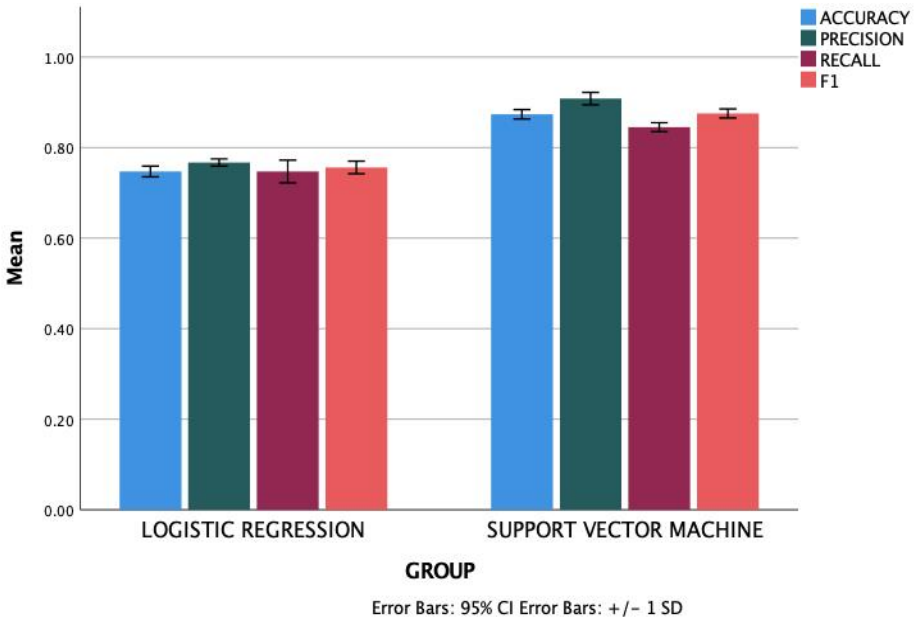

Fig. 1. Bar chart representing the comparison between Logistic Regression and Support Vector Machine algorithms in terms of mean precision,mean recall, mean accuracy, mean F1 for the prediction of cardiovascular diseases. Both the classifiers appear to produce similar rate accuracies but Support Vector Machine algorithms with slightly higher with an accuracy of 87.38%, the precision of 90.85%, recall of 84.52%, and F1 value of 87.55% when compared to Logistic Regression are with an accuracy of 74.73%, precision of 76.72%, recall value of 74.74%, and F1 value of 75.63%.Y-axis: Mean of accuracy, precision, recall rates for identification of keywords ± 1SD with 95% CI.

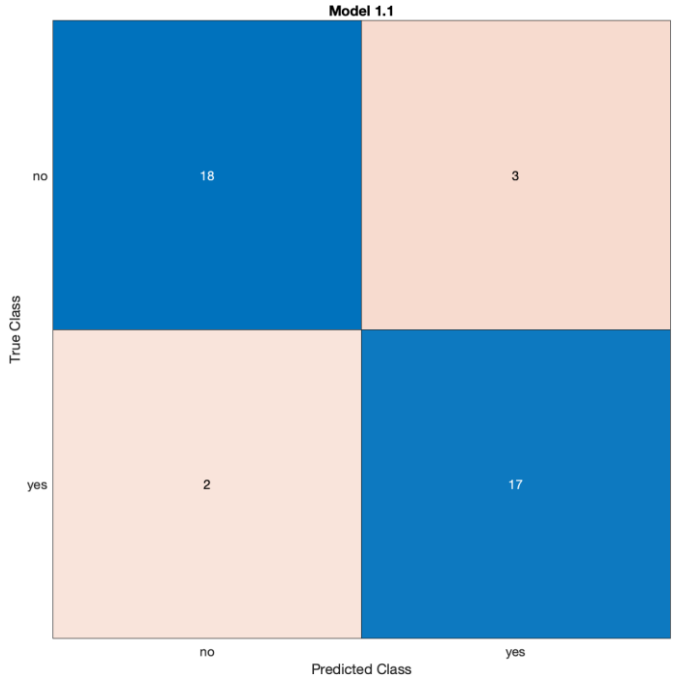

Fig. 2a. confusion matrix of Logistic Regression for K= 5

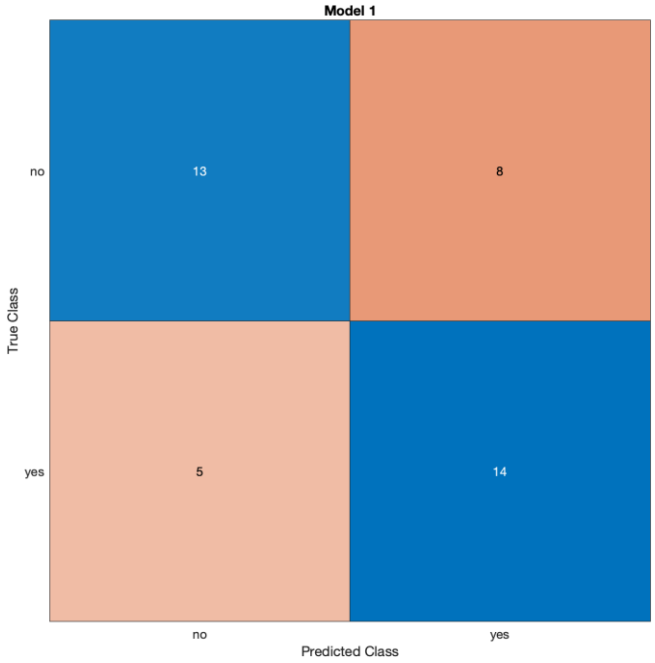

Fig. 2b. confusion matrix of Support Vector Machine for K=5