Analysis and Comparison of Random Forest Algorithm for Prediction of Cardiovascular Disease over Support Vector Machine Algorithm with Improved Precision

Rajvardhan Gadde, Neelam Sanjeev Kumar

Department of Biomedical Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamil Nadu, India, Pincode: 602105.

ABSTRACT

Aim: To find the best algorithm for the prediction of innovative cardiovascular disease accurately, with fewer errors between Random Forest and Support Vector Machine classifiers. Materials and Methods: Data collection containing various data points for predicting innovative cardiovascular disease from UCI machine learning repository. Classification is performed by Random Forest 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 Random Forest and Support Vector Machine. Support Vector Machine algorithm (87.38%) showed better results in comparison to Random Forest (83.50%). Conclusion: Support Vector Machine algorithm appears to give better accuracy than Random Forest for the prediction of innovative Cardiovascular Disease.

Keywords
Innovative Cardiovascular Disease, Machine Learning, Random Forest Algorithm, Support Vector Machine Algorithm, Accuracy, Precision

INTRODUCTION

The field of medical analysis has evolved drastically and it has raised the demand for the determination or prediction of the disease before the real disease occurs. These days, the death rate due to innovative cardiovascular diseases are increasing day by day every year. It is observed that 17.9 million deaths are being recorded as per the surveys made by WHO (Bharti et al. 2021). The way of living, new food habits and lack of physical exercise has been a major contributor for the increased innovative cardiovascular disease. For instance, smoking, high blood pressure, having high amounts of body fat, and consumption of junk are the main risk factors for innovative cardiovascular diseases mainly these can be suppressed in the early stages by changing the lifestyle, being more physically active, and taking medication in some complex cases (Das, Turkoglu, and Sengur 2009), (Ishaq et al. 2021). The main goal of this work is to apply machine learning algorithms to predict innovative cardiovascular diseases in this the comparison is done between Random Forest and support vector machine algorithms and finding the best accuracy obtaining algorithm (Al’Aref et al. 2019). The prediction of innovative cardiovascular disease should be done by the evaluation of certain features given by the user. If the accuracy of this prediction is high enough it will be helping to save a lot of human resources by avoiding incorrect diagnoses (Shouman, Turner, and Stocker 2012).

About 150 Science direct and 47 IEEE Xplore articles were found similar to this work in the last 5 years and has a clear report of developed algorithms and models using machine learning algorithms such as Naive Bayes, Support Vector Machine, Logistic Regression, Neural Network, Random Forest algorithms to predict and evaluate the performance of each algorithm in terms of accuracy, sensitivity, precision in the prediction of innovative cardiovascular diseases (Lai et al. 2019). In this paper, the major aim is to evaluate the validity of every algorithm in terms of accuracy, sensitivity, precision, and specificity and to find the best accuracy obtaining algorithm for the prediction of innovative cardiovascular disease.
Accuracy comparison is done over different classifiers Naive Bayes, Neural Network, Logistic Regression, Random Forest, diagnoses, and Support Vector Machine on UCI Machine Learning Repository data set. All these classifiers are executed in simulated environments using Matlab data mining tools. The executed results depict high accuracy by the Support Vector Machine algorithm with an accuracy of 87.38% and with the least error rate whereas the Random Forest algorithm got 83.50%. The precision values of the support vector machine and Random Forest algorithm also are 90.85% and 86.48% respectively, followed by recall and F1 values are also ruled out by the Support Vector Machine algorithm classifier with higher values than the Naive Bayes. Research work proposed a machine learning algorithm comparison of various classifiers to predict and reduce deaths due to innovative cardiovascular diseases (Dube et al. 2020). Our team has extensive knowledge and research experience that has translate into high quality publications (Chellapa et al. 2020; Lavanya, Kannan, and Arivalagan 2021; Raj R, D, and S 2020; Shilpa-Jain et al. 2021; S, R, and P 2021; Ramadoss, Padmanaban, and Subramanian 2022; Wu et al. 2020; Kalidoss, Umapathy, and Rani Thirunavukarasu 2021; Kaja et al. 2020; Antink et al. 2020; Paul et al. 2020; Malaikolundhan et al. 2020)

Inefficient predictions by the algorithms, more data redundancies, and human error in the early detection of innovative cardiovascular disease by conventional methods are giving errors this motivated me to do the research work improving the accuracy of the classifiers for the better prediction of innovative cardiovascular diseases. The major drawback of the existing research is having poor accuracy in the prediction algorithms. The authors are experts in machine learning and deep learning algorithms technologies. The main aim is to analyze and compare the breast cancer detection technologies using the best innovative machine learning algorithms namely the Random Forest algorithm and Support Vector Machine algorithm.

MATERIALS AND METHODS
The study was carried out at the University simulation laboratory, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai. In the current paper, the dataset was taken from UCI Machine Learning Repository innovative 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 innovative 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 (yangguang He et al. 2019) by clinical.com by keeping alpha error-threshold by 0.05, confidence interval at 95%, enrollment ratio as 0:1, and power at 80%. Sample preparation is carried out for Random Forest algorithm and Support Vector Machine algorithm for the data collected from the UCI Machine Learning Repository dataset.

The Random Forest algorithm is a probabilistic machine learning algorithm that is used for decision-making tasks (AhmedMedjahed, Saadi, and Benyettou 2013). Random Forest algorithm will approximate the independence between the features of the dataset rules. Applying the Random Forest algorithm model we can build a Naive Bayes Algorithm (xinlong Li et al.2019). All the characteristics in the Random Forest and this method is based on conditional probabilities. 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 the Random Forest algorithm with N value 20 and group 2 is Support Vector Machine algorithm with N value 20, the total sample size is 40. A sample dataset of both Random Forest algorithm and Support Vector Machine algorithm are exported to the Microsoft Excel Sheet for importing to Matlab as input. Matlab 2021a software has to be installed on the PC for training the source dataset. Both Random Forest algorithm and Support Vector Machine algorithm algorithms are used to train the sample groups. A confusion matrix is obtained and true positive, true negative, false positive, and false negative values are noted. accuracy (%), sensitivity (%) and precision (%)
values are calculated from the resulted 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 Random Forest algorithm with Support Vector Machine algorithm were done in this software. As the variables are independent of each other an independent sample T-test was carried out to find the 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 innovative cardiovascular diseases prediction by Random Forest and Support Vector Machine on UCI Machine Learning Repository, the results depict to produce the same variable results with the accuracy of 75.13% and 83.50%, precision 77.45% and 86.48%, recall 74.29% and 81.67% and F1 75.80% and 83.91% respectively. Table 2 shows the comparison of mean accuracy, mean sensitivity, and mean precision values of the Random Forest and Support Vector Machine. Support Vector Machine (SVM) algorithm. Support Vector Machine shows the higher values in terms of accuracy, sensitivity, and precision. Variable results with an accuracy rate of 87.38%, precision of 90.85%, recall value of 84.52%, and F1 value of 87.55%. Results of Random Forest are with an accuracy of 83.50%, pre-
cision of 86.48%, recall value of 81.67%, and F1 value of 75.80%. The Random Forest 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 Naive Bayes 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 representing the comparison of mean accuracy, mean precision, mean recall, and mean F1 values of Random Forest and Support Vector Machine as shown in Fig 1. Fig 2a and 2b represent the confusion matrix of Random Forest and Support vector machine respectively.

Table 3

<table>
<thead>
<tr>
<th>Accuracy</th>
<th>LEVENE'S TEST FOR EQUALITY OF VARIANCES</th>
<th>T-TEST FOR EQUALITY OF MEANS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F</td>
<td>SIG</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Equal Variance Assumed</td>
<td>2.80</td>
<td>.102</td>
</tr>
<tr>
<td>Equal Variance is not Assumed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Precision</td>
<td>6.74</td>
<td>.013</td>
</tr>
<tr>
<td>Equal Variance Assumed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Equal Variance is not Assumed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Recall</td>
<td>.147</td>
<td>.704</td>
</tr>
<tr>
<td>Equal Variance Assumed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Equal Variance is not Assumed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>F1</td>
<td>.884</td>
<td>.353</td>
</tr>
<tr>
<td>Equal Variance Assumed</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Fig. 1 Bar chart representing the comparison between Random Forest and Support Vector Machine algorithms in terms of mean accuracy, mean precision, mean recall, 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 accuracy, precision, recall, and F1 values of 87.38%, 90.85%, 84.52%, and 87.55% respectively. Random Forest are with an accuracy of 83.50%, precision of 86.48%, recall value of 81.67%, and F1 value of 75.80%. Y-axis: Mean of accuracy, precision, recall rates for identification of keywords ± 1SD with 95% CI.
DISCUSSION

In this research paper for the prediction of innovative cardiovascular diseases, we observed 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 Random Forest are with an accuracy of 83.50%, precision of 86.48%, recall value of 81.67%, and F1 value of 75.80%. Although not statistically significant, the significant difference appears to have slightly increased table 3. Machine Learning algorithms play an important role in the early detection of innovative cardiovascular diseases.

Related works are done by many researchers (Torres 2021) proposed using similar comparison and by using machine learning algorithms and the main aim is to accurately evaluate the model in terms of accuracy, precision, sensitivity, specificity, and F-measure. Another study was done by (Haq et al. 2018) this paper the author implemented a Machine Learning algorithm for the prediction of innovative cardiovascular diseases and by using a cardiovascular dataset which resulted from Accuracy, sensitivity, specificity, and MCC. A paper by (Nikam et al. 2020) used a similar feature section by using similar machine learning algorithms in which Naive Bayes had shown a lower accuracy value of 83.50% for the prediction of innovative cardiovascular disease. 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 accuracy, precision, recall, and F1 values. 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 machine learning algorithms and deep learning algorithms 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, the support vector machine algorithm has a higher accuracy of 87.38% than the Random Forest which has an accuracy of 83.50%. 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 Random Forest are with an accuracy of 83.50%, precision of 86.48%, recall value of 81.67%, and F1 value of 75.80%. 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’s learning-Based guide 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