Classification and Prediction of Heart Disease using Novel Random Forest Algorithm by Comparing Logistic Regression for Obtaining Better Accuracy

T Poojitha, Mahaveerakannan R

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:** Heart attacks are usually caused due to blockages, partially or completely, of the heart's veins or arteries that constrict the flow of blood from or to the heart. The primary objective of this review aims to be seen as the most appropriate algorithm to give us the ideal prediction. We will be comparing the novel Random Forest with Logistic regression to find out which of these can give us the best accuracy.

**Material and Methods:** The study used 143 samples with novel Random Forest and Logistic Regression is executed with varying training and testing splits for foreseeing the accuracy of coronary disease prediction with the 80% of G-power value and heart disease data were gathered from multiple web sources, including latest The study's findings and criterion were 0.05%, with a 95% probability value, average, and confidence interval. The performance accuracy rate of the classifiers is used to evaluate the coronary disease dataset. There was a statistically significant value test between the novel Random Forest and Logistic Regression is 0.046 (p<0.05).

**Results and Discussion:** The accuracy of predicting coronary disease in the novel Random Forest 90.16 % and Logistic Regression 85.25 % is obtained. **Conclusion:** This study concludes that the Prediction of Coronary disease using the novel Random Forest (RF) algorithm looks to be fundamentally superior to the Logistic Regression (LR) with increased precision.

**Keywords**

Coronary Disease, Novel Random Forest Algorithm, Logistic Regression, Machine Learning, Classification, Prediction.

**INTRODUCTION:**

The prediction of Coronary Disease and to increase the precision of predicting the disease using novel Random Forest and comparing it with Logistic Regression. In the course of the last 10 years, coronary disease or cardiovascular sickness remains the essential premise of death around the world (Steiner et al. 2021). An estimation by means of the World wellbeing association, that over 17.9 million passouts happen each overall year due to cardiovascular disorders, mainly 80% of the passouts are only in view of coronary vein infection also, cerebral stroke (Woodruff et al. 2021). Estimating an individual's chances of adequate blood flow to the heart is one of the significant issues of cardiovascular disease. The prediction technique might be built by applying a number of longitudinal review study autonomous-regression analyses (Sharma 2021). The main motivation of this paper is to focus on diagnostic tests and see some of the daily health habits that contribute to cardiovascular disease(CVD). Most importantly, the different machine learning algorithms are compared using intelligent optimization algorithms. Manual classification data is used for healthy or unhealthy (Tithi et al. 2019). Data science plays a fundamental role in handling a gigantic information measurement in the sphere of medical services Because cardiovascular disease is a perplexing task, there is a need to automate the schedule phase in minimizing any problems and to advise the patient far ahead of time (Rajdhan et al. 2020a). Finding and predicting Cardiovascular infection (CVD) requires more accuracy, flawlessness, and rightness on the grounds that a little mix-up can cause weariness issues or the death of the individual. The significance of this exploratory study is to bring coronary illness forecasts utilizing AI that can assist with recognizing the clinical information, for example, hypertension, diabetes, cigarettes smoked each day, etc is taken as information and afterward, these elements are modeled for prediction(Maranhao Neto et al. 2021). At the same time, we can apply this research in clinical field correctness and in some complicated
tasks like fatigue problems. 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; Sudha S, R, and P 2021; Ramadoss, Padmanaban, and Subramanian 2022; Wu et al. 2020; Kailidoss, Umapathy, and Rani Thirunavukkarasu 2021; Kaja et al. 2020; Antink et al. 2020; Paul et al. 2020; Malaikolundhan et al. 2020)

A ton of examinations has been finished based on the classification of Hearts utilizing Machine Learning Algorithms. More than 120 articles were published in Science Direct and nearly 140 articles were found in google scholar. Related to work on coronary disease prediction using machine learning the authors obtained better and more accurate results with the application of KNN accuracy of 97.3% (Nambi 2018). Coronary illness prediction and classification utilizing AI calculations enhanced by molecule swarm streamlining and subterranean insect province improvement, the authors achieved better and more accurate results having 99.65% (J. Singh 2017). Comparing different algorithms of novel Random Forest arrangement looking for better execution in coronary illness determination, WEKA (R. Singh and Rajesh 2019), classification system using Logistic model tree algorithm and Novel Random Forest Algorithm, the authors gained accurate results using different techniques (Wang et al. 2021). A half and half savvy framework system for the forecast of coronary illness utilizing AI calculations (Westerlund et al. 2021), the intrusive based methods to the diagnosing of the coronary illness depend on concerning symptoms to address these intricacies painless clinical choice emotionally supportive network comprises of the fake neural organization (ANN), Logistic relapse (LR), AdaBoost (AB), support vector machine (SVM), Fuzzy Logic (FL), Naive Bayes (NB), and unpleasant set [9,10] has been created has been created by the different researches and broadly utilized for coronary illness conclusion (Baig 2020).

The experimental results indicate that the novel Random Forest yields 89.14% in terms of the accuracy for Resting ECG and CP (chest pain) of the datasets. This scientific study has been presented and published in over 55 indexed journals. (Rajdhan et al. 2020b).

The existing research study predicts the classification using probability but it does not group the attributes (Qi and Zhang, n.d.). Another research paper has developed the performance of the order and its dataset will yield a limit of 54.30% of precision utilizing AI calculations and this will make an exploration hole to foresee the reliable and sensible ways of managing making an early finding to achieve brief organization of the infection (Chatterjee and Chandran 2019). The aim of this research is to utilize the original Random Forest (RF) algorithm to group the best arrangement of value administrations and compelling determination of anticipating coronary illness at an early stage by achieving better accuracy.

**MATERIALS AND METHODS:**

This exploration study was done in the Artificial Intelligence Research center at the Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai. The innovative Random Forest (RF) method and the Logistic Regression (LR) technique are used in this research project. The G-power programme version 3.1.10 was used to forecast each scale factor, yielding 286 total number of observations and 143 number of respondents per group with 80% G power values, a minimum significant value of 0.05, and a probability value of 95%. (Rajdhan et al. 2020b).

This research makes use of a dataset obtained from the kaggle open-access dataset. One of the most popular online networks for information researchers and artificial intelligence professionals. It allows clients to search for and discover the datasets they need, as well as give a customisable Jupyter notebook environment with free GPU and 6 hours of runtime. The dataset consists of 76 attributes (Westerlund et al. 2021). The attributes are age, sex, chest torment type (4 sorts), resting circulatory strain, serum cholesterol in mg/dl, fasting glucose > 120 mg/dl, most extreme pulses accomplished, and 6 others. The dataset has 14 columns and 304 rows containing data on symptoms associated with coronary disease, including duplicate, null, and missing values. Pre-processing of the datasets is done to remove the duplicate and null values with the help of functions that are provided by Microsoft excel. The dataset that is collected consists of all numerical forms; there is no need to convert the data. The unwanted attribute such as age, exacting attributes that are not necessary for the prediction is removed (Saranya S et al. 2020).

RF falls within the umbrella of adaptive boosting. RFs may be utilized for effectiveness as well as foresight. For implementing an RF, the random state function is used to create forest-like structures and com-
bine them. The overfitting of data is done with help of calculating lowered, this is accomplished by dividing the dataset and assessing all splits by creating the rf. Predict.

The RF method uses the following pseudocode to apply a coronary disease dataset. The datasets will be fed into the pseudocode, and the final result will be passed through the accuracy and classification parameters.

Pseudocode of the Novel Random Forest Algorithm

```python
from sklearn.ensemble import RandomForestClassifier

max_accuracy = 0

for x in range(2000):
    rf = RandomForestClassifier(random_state=x)
    rf.fit(X_train, Y_train)
    Y_pred_rf = rf.predict(X_test)
    current_accuracy = round((accuracy_score(Y_pred_rf, y_test) * 100, 2)
    if (current_accuracy > max_accuracy):
        max_accuracy = current_accuracy
        best_x = x

print(max_accuracy)
print(best_x)
```

Innovative Random Forest Algorithm is extensively used for both recurrence and identification issues that must be addressed. This algorithm can be used in various places such as banking, prediction works, health, stock markets, artificial intelligence, etc. To get more accurate and stable estimates the novel random forest will create a forest-like structure and combine them. Subsets from both datasets and attributes are selected arbitrarily and get trained. using this method overfitting of data can be lowered. This algorithm takes lower training time than many other algorithms on large datasets with maintaining the precision of the accuracy when a huge part of data is not present.

Pseudocode of the Logistic Regression algorithm

```python
from sklearn.linear_model import LogisticRegression

lr = LogisticRegression()

lr.fit(X_train, Y_train)

Y_pred_lr = lr.predict(X_test)
```

Logistic regression is essentially an administered grouping computation. In a representation problem, the objective variable (or yield), y, might have essentially distinct features for a particular situation approach of features (or inputs), X. Contrary to standard reasoning, vital backslide is a backslide model. The model constructs a backslide technique to forecast the possibility that a given data segment will have a place in the class labelled «1.» When a decision edge is introduced, it converts into a request methodology. The edge's positioning regard is an indispensable piece of Logistic backslide and is dependent upon the course of action issue itself. The decision for the value of the edge regard is fundamentally affected by the potential gains of precision and survey.

Statistical Analysis

The statistical analysis is applied using For each group, descriptive statistical analysis such as mean, standard deviation, and standard error were performed using the Statistical Package for Social Sciences (SPSS). The test was performed to analyze factors across the review gatherings. Hence the independent variable in this study is mean, entropy, variance, and contrast (Woodruff et al. 2021). The dependent variables are fever, heart rate, bp, cholesterol. The independent sample t-test is performed to compare the accuracy performance of both Group-1 and Group-2 using the RF and the LR algorithms.

RESULTS

The RF and LR learning algorithms are compared with 10 examples using various 70% training and 30% evaluation samples with varied amounts of entries in the dataset, and the conclusions are displayed in Table 1. The dataset contains 304 rows, and the accuracy of the RF and LR algorithms is determined for 10 samples (iterations). It is observed that the RF algorithm is significantly better than the LR algorithm. The preprocessing is used to clean the data to apply for the machine learning algorithm, due to the original dataset containing more errors and null values, it has been preprocessed. The dataset has been resized and made ready for the next stage, as the network allows. The algorithm for preprocessing the dataset can be implemented using Spyder functions.

From Table 1, the statistical analysis of the Novel RF algorithm and LR algorithm. For 10 iterations, the mean accuracy value, standard deviation, and accuracy for Novel RF and LR algorithms are obtained. The RF algorithm performed 90.16% better than the LR method (85.25%).
In Table 2 we observed after performing statistical analysis of the Novel RF algorithm and the LR algorithm. Mean accuracy value, the standard deviation, and standard error mean are calculated after 10 iterations. The mean of the RF 89.14% algorithm fared better than the LR 84.12% method.

<table>
<thead>
<tr>
<th>Algorithms (Accuracy)</th>
<th>Sample (N)</th>
<th>Mean</th>
<th>Std Deviation</th>
<th>Std error mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Forest</td>
<td>10</td>
<td>89.14</td>
<td>2.9654</td>
<td>0.93775</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>10</td>
<td>84.12</td>
<td>3.7039</td>
<td>0.17128</td>
</tr>
</tbody>
</table>

In Table 2 we observed after performing statistical analysis of the Novel RF algorithm and the LR algorithm. Mean accuracy value, the standard deviation is observed that the RF algorithm performed better than the LR algorithm. The Novel RF algorithm obtained 2.9654 standard deviations with 0.93775 standard error while the LR algorithm obtained 3.7039 standard deviations with 0.17128 standard error. After comparing the 10 accuracies of both the algorithms, a random forest algorithm was obtained with the mean of 89.14% and logistic regression obtained with the mean of 84.12. Also, the independent sample t-test was utilized to look at the exactness of two calculations and a genuinely tremendous distinction was seen with the RF model having 90.16% precision and LR has 85.25% of exactness. When contrasted with the other calculation’s exhibition, the proposed RF classifier is fundamentally better compared to the SVM algorithm.

Table 3 shows the separate sample t-test with relevance values and standard error. The «p» value was less than 0.05, which is deemed clinically meaningful, and 95% confidence intervals were produced. The independent sample t-test was utilized to think about the precision of two calculations and a genuinely massive distinction was taken, note is 0.046 (p<0.05) The Novel RF model obtained 90.16% accuracy. When compared to the performance of the other strategies, the suggested Novel RF outperforms LR substantially.

The symptom shown here is the remainder of the ECG comparison as seen in Fig. The red individuals are at danger because of this one attribute, whereas the blue people are not at risk because of this one attribute; the values range from 0 to 300.

From Figure 2 Comparison of the chest pain between people of different age groups having the pain
in the heart using target attribute, the values range from 0.0 to 0.8.

From Figure 3 Comparison of attributes between thal and density to determine the chest pain intensity the values range between 0.00 to 2.00.

From Figure 4 the mean accuracy of the two algorithms has been shown and compared the mean error and the graph shows that Novel RF is better than the LR algorithm.

The Novel RF model obtained 90.16% and the LR algorithm already predicted 85.25% performance value. Finally, as shown in Fig. 4, we compared the mean error of the Novel RF and LR algorithms. As a result, the Novel RF method outperforms the LR algorithm substantially.

DISCUSSION

Novel random forest as convincingly appears better than Logistic Regression with improved accuracy. The novel Random Forest classifier shows some When compared to Logistic Regression, There is a distinction in terms of effectiveness, scoring rate, and overall efficiency. Data mining is crucial in the healthcare business for new illness identification. A variety of tests must be performed on the patient in order to diagnose an ailment. On the other hand, using data mining techniques can minimize the number of tests necessary.

To Support this work, a Comparison between each algorithm’s performance in terms of accuracy, and loss proved that Random Forest is the best overall. The
outcome of this paper is similar to (Shah, Patel, and Bharti 2020) where they found that the Novel Random Forest Algorithm shows better performance with 89% accuracy than Logistic Regression. Hence the novel Random Forest can be used to predict heart disease efficiently (Zhang et al. 2021). The Positive other finding backup, the experiment was carried out using regressive testing of all possible variations that are there for the given models (Risk Prediction in Cardiovascular Disease-Current Status and Future Challenges 2009; Kaur, Sharma, and Sharma 2019). Other study findings, in contrast to this study, the novel Random Forest method involved the use of various tree depth levels along with the use of confidence vote and Gini index criteria. It resulted in an accuracy of 89.14% (Risk Prediction in Cardiovascular Disease-Current Status and Future Challenges 2009; Kaur, Sharma, and Sharma 2019)

Some other research findings employed a low sample size to counteract this study number of algorithms for classification but we are using as many as possible machine learning algorithms for analysis. In other existing studies, they only applied for specific calculations advanced by molecule multitude and subterranean insect state improvement (Saranya S et al. 2020). But In this proposed research work, it can be found that the staging framework has an enormous number of hyperparameters that are to be calibrated. The classification models can be enhanced in order to improve the features of classification (Tithi et al. 2019).

Most of the previous work done is based on the data from the specific dataset. The limitation of this research study having a lack of accuracy in predicting heart disease is improved by machine learning algorithms for analysis. In the future, more attributes in the dataset can be included and can also be used as datasets for better accuracy. Artificial intelligence can also be utilized in the future to anticipate the seriousness of coronary illness.

CONCLUSION

The main aim of the study is to quantify the precision in the classification of heart disease using medical data. This research study applied RF algorithm for the detection of heart disease from a dataset with parameters achieved has been compared with the LR algorithm. The results obtained show that the RF algorithm has found 90.16% of accuracy on predicting heart disease than the 85.25% of the LR algorithm.

DECLARATION

Conflict of Interests

No conflict of interest in this manuscript.

Authors’ Contributions

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

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 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. Verk Technologies, Rangareddy, Telangana
2. Saveetha University
3. Saveetha Institute of Medical And Technical Sciences
4. Saveetha School of Engineering

REFERENCES


28. Wang, Ke, Jing Tian, Chu Zheng, Hong Yang, Jia Ren, Yanling Liu, Qinghua Han, and Yanbo Zhang. 2021. “Interpretable Prediction of 3-Year All-Cause Mortality in Patients with Heart Failure Caused by Coronary Heart Disease Based on Machine Learning and SHAP.” Computers in Biology and Medicine 137 (October): 104813.