**Heat Disease Prediction: A Comparative Study of Random Forest and Logistic Regression**

**Bhagyalakshmi Katta, Saipreetham Kadara, Keerthi Kalakuntla, Ganesh Babu Athota**

**ABSTRACT**

Heart disease is the main reason of sickness and fatalities worldwide. Early detection and accurate risk assessment can stop heart disease before it starts. Recently, heart disease risk prediction using ML algorithms has produced encouraging results. One such ML algorithm available for heart disease prediction has been utilized in various research is Random Forest. In this study, we used a dataset that contained data on a variety of patient demographic, clinical, and lifestyle characteristics and applied the Random Forest algorithm and the Logistic Regression method to it. The dataset contains information about age, gender, blood pressure, cholesterol levels, resting ECG, thalamus, Ex angulation, and other variables. This data were used to train Random Forest and Logistic Regression models to predict the existence of heart disease. We examined a few factors, including F1 score, Accuracy ,Recall and Precision to validate how our designed model is working. In comparison to logistic regression, the findings demonstrated that the Random Forest algorithm had a high accuracy of 95% in predicting the risk of heart disease. Our study suggests that the Random Forest algorithm may be a valuable tool for evaluating the risk of acquiring heart disease and for identifying high-risk individuals who might benefit from early interventions and preventive measures.

**INTRODUCTION**

Heart disease is the primary cause of global mortality, causing up to 1.79 million deaths annually, as reported by the World Health Organization (WHO) in 2021. Machine learning techniques consider various risk factors such as diabetes, cholesterol levels, abnormal pulse rate and ECG can help forecast heart attacks [1]. In a 2020 study, wearable mobile technology data was utilized to predict the likelihood of heart disease using Naive Bayes, SVM, and functional tree algorithms. The K-Nearest Neighbor model had the 90.8% with highest accuracy whereas other ML algorithms had less accuracy of 80.3% [2]. Earlier studies have also investigated the application of machine learning approaches for cardiac disease diagnosis and categorization. Rather than enhancing current methods with optimized strategies, this research concentrates on the distinct effects of various machine learning techniques. The use of hybrid optimization methodologies for machine learning classification optimization is uncommon in academia [3].ML algorithms examine huge amounts of intricate and diverse data, including lifestyle factors, medical histories, demographic details, to detect patterns and generate predictions [4]. The high incidence and fatality rates of heart disease have stimulated interest in using machine learning to forecast its occurrence. Early detection and treatment can substantially enhance patient outcomes [5].

In this paper, we compare the Random Forest Algorithm and Logistic Regression for heart disease prediction. The following is how the paper is organized: we outlined our study methodology in section2 . In Section 3, we present our research findings and analyze the results. Finally, in section 4 we summarize and conclude the paper.

**METHODOLOGY**

According to the WHO, heart disease causes 17.9 million deaths per year, making it one of the main causes of fatalities [6]. With the aid of early detection, people can change their habits and, if necessary, ensure that they receive top-notch medical care. ML can be used to streamline and minimize the symptoms of heart disease. Our study's objective is to create a machine-learning model that accurately predicts a patient's risk of getting heart disease based on their medical background and other pertinent information. We decided to use the random forest technique and logistic regression to solve this issue since research has demonstrated that the random forest algorithm and logistic regression are the best algorithms for predicting cardiac disease [7]. Random forests and logistic regression, which provide the maximum accuracy, are the best prediction methods. Large datasets with thousands of variables can also be managed using this technique [8]. Overall, using machine learning to forecast heart disease may improve patient outcomes and reduce the strain on healthcare systems. Fig:1 is a flow chart which describes the whole process involved in the study from collection of data to training the models.

Graphical user interface, diagram

Description automatically generated

Fig:1 Flowchart describing the steps involved in the methodology.

**Data Collection**:

For our research, we utilized the pre-processed UCI Cleveland dataset from the Kaggle and open-source websites [9]. 304 data instances and 14 attributes were merged to forecast heart disease. After the gathering of many records, the data connected to heart disease is pre-processed and analysed. Fig 2: is a graph showing the patients presence of heart disease and no heart diseases considered for our study.

Chart, bar chart

Description automatically generated

Fig 2: Graph depicting the patients with heart disease and without heart disease.

**Data Pre-Processing:**

Pre-processing of data typically entails feature engineering, data cleansing, and formatting [10]. There were no duplicates or missing data in the 304 patient records that make up the pre-processed dataset we used. The results show that 138 records have a 0 value, which indicates no heart illness, whereas 165 records have a 1 value, which shows the presence of heart disease. The dataset was also split between training and testing sets using a 70:30 split.

Chart, bar chart

Description automatically generated

Fig 3: Graph depicting the frequency of heart disease in different genders.

Chart, bar chart

Description automatically generated

Fig 4: Line graph showing the heart diseases for different ages.

**Data Analysis:**

The characteristics that were shown to be most crucial for predicting heart disease were chosen using a correlation matrix. Age, sex, the kind of chest pain, blood pressure, cholesterol, heart rate, and exercise-induced angina were determined to be the most crucial factors. Fig 3 is the graph showing the presence and absence of heart diseases in different genders considered for our study. Random Forest has been selected as the ML method for predicting the heart diseases. Fig 4 depicts the heart disease frequency in different age groups considered for the study. This technique was chosen because it is frequently employed for binary classification tasks and because its implementation is quite straightforward [11]. The system was tested using measures including accuracy, precision, and recall after being trained on the training set.

**Model Evaluation:**

Accuracy, precision, and recall were the evaluation metrics employed in our study.

**RESULTS AND DISCUSSIONS**

ML Algorithms such as Logistic Regression and Random Forest are frequently utilized for classification tasks. For instance, when predicting the likelihood of heart disease, these algorithms can identify whether a person is at risk based on specific characteristics or risk factors. The accuracy Logistic Regression and Random Forest in predicting heart disease is caused by multiple factors, such as the dataset's size and quality, the feature selection, and the tuning of the algorithm's hyperparameters. Overall, these machine learning techniques are potent tools that can effectively forecast the presence of heart disease based on certain risk factors. Random Forest and Logistic Regression were two machine learning models whose performance in foretelling the absence or presence of heart disease was assessed. The Random Forest model received an accuracy score of 95.0%, correctly predicting 95.0% of cases in which heart disease was present or absent. It had a precision score of 90.0%, meaning that it correctly recognized 90.0% of the positive instances (patients with heart disease). In terms of recall, the model correctly identified 89.0% of the positive cases (patients with heart illness). On the other hand, the logistic regression model achieved an accuracy score of 90.0% and a precision score of 90.0%. It correctly recognized 90.0% of the positive instances (patients with heart disease) and effectively identified 91.0% of the positive cases. Overall, both models showed good performance in predicting heart disease, but the random forest model performed slightly better in terms of accuracy and recall, while the logistic regression model had a slightly higher precision.

**CONCLUSION**

The aim of the study was to compare the accuracy of heart disease prediction using the two ML algorithms namely Logistic Regression and Random Forest. These two algorithms were used with the dataset we downloaded from Kaggle to create the prediction models. Later, the performance of the two models was compared using factors like F1 score, accuracy, Recall and Precision.

The accuracy of the Random Forest method is 95%, while that of Logistic Regression is 90%. The results of the investigation showed that in every measure, the random forest algorithm outperformed the logistic regression method.

In conclusion we can say that Random Forest should be used to forecast heart disease rather than logistic regression. One drawback of our study could be that we used only one data set for the study. We are also interested to see how well our study of various ML Algorithms like Decision Trees and SVM’s, predict heart disease in the future. Additionally, we suggest including data from various sources and combining them to validate the study's findings and build more accurate predictive models.

**REFERENCES**

1. Saravanan, K., Priya, R. S., & Vignesh, M. (2021). Machine learning for heart disease prediction: a review. Journal of Medical Systems, 45(6), 1-16.
2. Islam, M. S., Ahmed, S. I., Hasan, M. T., Siddiqui, S. A., & Islam, M. A. (2020). Performance comparison of four machine learning algorithms in heart disease prediction. SN Computer Science, 1(5), 1-12.
3. Gaurav, K., Sharma, A., & Bhatia, G. (2021). A review of machine learning techniques for heart disease prediction. International Journal of Information Technology and Management, 20(1-2), 51-67.
4. Sharma, M., & Tripathi, R. K. (2020). Intelligent prediction of heart disease using machine learning techniques. Journal of Ambient Intelligence and Humanized Computing, 11(8), 3293-3305.
5. Dong, X., Li, Y., Zhu, Y., Li, H., & Li, D. (2020). Intelligent prediction of heart disease based on machine learning. BMC Medical Informatics and Decision Making, 20(1), 1-12.
6. Apurv Garg et al 2021 IOP Conf. Ser.: Mater. Sci. Eng. **1022** 012046
7. Victor Chang, Vallabhanent Rupa Bhavani, Ariel Qianwen Xu, MA Hossain, An artificial intelligence model for heart disease detection using machine learning algorithms, Healthcare Analytics,Volume 2 2022 100016.
8. Shah, Devansh & Patel, Samir & Bharti, Drsantosh. (2020). Heart Disease Prediction using Machine Learning Techniques. SN Computer Science. 1. 10.1007/s42979-020-00365-y.
9. S. Mohan, C. Thirumalai and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," in IEEE Access, vol. 7, pp. 81542-81554, 2019, doi: 10.1109/ACCESS.2019.2923707.
10. Apurb Rajdhan , Avi Agarwal , Milan Sai , Dundigalla Ravi, Dr. Poonam Ghuli, 2020, Heart Disease Prediction using Machine Learning, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 09, Issue 04 (April 2020),
11. N. M. Lutimath, N. Sharma and B. K. Byregowda, "Prediction of Heart Disease using Random Forest," 2021 Emerging Trends in Industry 4.0 (ETI 4.0), Raigarh, India, 2021, pp. 1-4, doi: 10.1109/ETI4.051663.2021.9619208.