Vol. 2 No. 3 (2024)

Spectrum of Engineering Sciences

Online ISSN
3007-3138
Print ISSN
3007-312X

Predicting Early Heart Disease: A Supervised Machine Learning Approaches

Hira Latif
Department of Computer Science, University of Agriculture, Faisalabad. hiralatifg@gmail.com
Tayub Shaheen
University of Lahore, Software Engineering Department
hafiztayubshaheen@gmail.com
Habib ur Rehman
Software Engineering Department, University of Lahore
habib.Rehman@SE.uol.edu.pk
Awais Rasool
Software Engineering Department, University of Lahore
awais.rasool@se.uol.edu.pk

Abstract

Over the past few years, worldwide, "heart-related" illnesses are now the main cause of death. The primary focus of this study is that it raises a patient's risk of cardiovascular disease (CVD) based on a variety of medical variables. Manual methods for diagnosing heart disease are inaccurate a key requirement for solving the issue is the development of a disease awareness prediction system. To identify and classify people with heart infection, machine learning processes are effective. Machine learning techniques have recently been used to help the medical field and professionals identify heart-related diseases. To diagnose cardiac disorders, by using algorithm supervised learning to build a mathematical model from a set of information that includes the necessary inputs and outputs. The supervised learning techniques including Random Forest,

logistic regression, naive Bayes, K Nearest Neighbor, and decision trees are employed. This research aims to provide doctors more confidence and accuracy in their predictions by utilizing actual data from patients, both well and sick. Parameters use in cardiovascular disease focus on Sex, Sugar level, Blood pressure, Cholesterol, and Hypertension. Python language is used with the help of google colab tool for the implementation of these algorithms. Main reason of this research is that to increase the accuracy rate. This research gives accuracy 86.89% by using Random Forest Algorithms with the help of less parameter. This accuracy indicates a reasonably accurate model, it's also important to evaluate the performance in the context of the specific problem you are trying to solve.
Keywords: Machine Learning, Supervised Learning, Cardiovascular disease, Accuracy

1. Introduction

The heart is the most essential organ in the human body as it serves as the powerhouse of the circulatory system that regulates blood flow. "Cardiovascular disease" (CVD) includes a range of illnesses and disorders affecting this system (Pal and Parija, 2021). Coronary artery disease is the cause of most deaths with about 2% annual growth, affecting over 26 million people. CAD had claimed 17.5 million deaths in the year 2005, while the present estimate indicates that nearly 2% of the population suffers from this disease, with 10% of them more than 65 years old. It has projected that CVDs will cause approximately 23.6 million extra deaths in 2030. In Pakistan, though much achievement has been made in controlling infectious diseases, CVDs, diabetes, and cancer contribute most to early mortality. In Pakistan, even though faced with all these problems, life expectancy went up from 61.1 to 65.9 years in the last three decades.

Heart disease is prevalent among both men and women from different parts of the world; therefore it is logical that risk factors

should be taken into account during taking any decisions or lifestyle choices. Although the influence of hereditary factors is tremendous, lifestyle plays a significant role in affecting heart problems. The most crucial risk factors are age, gender, and a family history while also incorporating smoking as a behavior. Malnutrition, hypertension, cholesterol accumulation, diabetes, obesity, inactivity, stress, and poor personal cleanliness are significant concerns (Patro *et al.*, 2021). Diagnosing heart disease can be challenging and complicated. It is based on the scrutiny of signs and symptoms that are accompanied by a proper physical examination. Factors that can affect diagnosis are false premises and unpredictable findings (P. U. Anitha *et al.*, 2022). The good news is that cardiovascular disease can be prevented if risk factors, among which are an unhealthy diet, a sedentary lifestyle, and the use of tobacco and alcohol, are addressed. This means people at high risk are those with chest pain, those with high cholesterol, diabetes mellitus, and hypertension, who need early detection and prediction mechanisms to be able to manage their health and prevent sudden heart failure (Tougui et al., 2020).

Despite the many programs aimed at preventing disease, several issues remain. Two of the significant areas of concern in heart disease research are: (1) obtaining relevant features from datasets for heart patients, as mainly the features derived may not be relevant, and (2) development of efficient prediction systems that would cope with high volumes of data in real-time time. The study would ensure a low number of deaths through the early diagnosis of cardiac problems using a new model that emphasizes the link between cardiovascular factors and other health elements. Other than that, the research was comparing various classification algorithms in the detection of cardiovascular disease, providing useful knowledge to many (Singh and Kumar, 2020).Machine learning (ML) makes the machine process the information more

efficiently, especially when insights from them seem hard to gain. As the datasets are available on a larger scale, ML is needed in each of their domains. Basically, the core goal of machine learning refers to the realization of machines in which inputs of data are effectively learned, and many studies have centered on the development of algorithms that will give conditions under which machines can learn independently without explicit programming. Many mathematicians and programmers explore diverse approaches for problems associated with large dataset (Mahesh, 2020).

This paper focuses on different ML classification approaches that assist health professionals in the proper diagnosis of cardiovascular conditions. Using algorithms that weigh and crunch big data, ML can improve the speed and accuracy of diagnosis hence the quality of care a patient receives (Diwakar *et al.*, 2020). In machine learning, both supervised and unsupervised learning are the two primary categories. Unsupervised learning works with data that only contains inputs and no output labels, while supervised learning algorithms build mathematical models from datasets that have both inputs and outputs (Jiang, 2020). This work is based on supervised machine learning techniques, such as KNN, SVM, DT, NB, LR and RF for detecting cardiac problems. These algorithms predict the severity of disease at early stages, which can act as a reliable source of assistance for medical personnel. This research is trying to improve the predictive algorithms of heart disease. It hopes that improved accuracy in diagnosis leads to better early diagnosis of cardiovascular diseases. Finally, it wants to contribute toward reduction of heart disease mortality. This approach would improve patient outcomes and, by extension, public health to unprecedented levels.

2. Related Work

Tarawneh and Embarak (2019) applied hybrid data mining classifiers on 303 records of the UCI repository to predict heart disease, with 85.55% accuracy after reducing features from 14 to 12. He has used six algorithms: KNN, NN, SVM, J48, RF, and NB. Training and testing dataset split into 70 and 30. Kompella et al. (2019) used MLP-NN for cardiac disease prediction. It achieved a fitness value of 83.39% for three different cases in 3.86 seconds based on data from 303 cases in the Cleveland machine learning dataset. The preprocessing included all six files to remove missing values, and it focused on 14 variables of heart disease. Saleh Alotaibi (2019) system based on a machine learning approach would be a great requirement in order to classify data and bring out rules.

Such research is, therefore, proposing and implementing a model using a combination of five different algorithms. The model was developed with Rapid Miner, and the outcomes were proved to have more accuracy than Matlab and Weka. Sujatha and Mahalakshmi (2020) The classification of cardiac illness is done in this study using a variety of methods, such as Decision Tree, Naïve Bayes, Random Forest, SVM, KNN, and Logistic Regression. The Random Forest has the highest accuracy of 83.52% among them. (Shah *et al.*, 2020) this research validates the heart disease attributes using supervised learning algorithms, contains the Cleveland database's 303 instances of K-nearest neighbor (KNN).

From the results obtained, KNN showed better predictive accuracy compared with the rest. (Angraal *et al.*, 2020) focus to the rising incidence of heart disease and the urgent need for a diagnosis using data mining and machine learning approaches. They focus on 14 important characteristics based on research of the Cleveland database, and the K-nearest Neighbour method provides the most accurate prediction of heart disease risk. Katarya

and Meena (2021) also said that busy lifestyles lead to unattended health conditions, and growing heart disease, accounting for more than 31% of deaths globally has been stated by WHO data. The study aims to identify the causes related to heart syndrome with the help of several machine learning methodologies for predictive purposes.

Ansarullah *et al.* (2022) a non-invasive heart disease risk assessment model is developed using weighted attributes chosen by cardiologists. Among various classifiers, it performed well with better predictive metrics discovered, especially with the random forest model. The model is most helpful in areas that do not have early primary medical care with advanced technology to detect heart diseases. (Bhavekar and Goswami, 2022) address the requirement of non-invasive, cost-effective early diagnosis of cardiovascular disease through hybrid deep learning approach using RNN and LSTM techniques this hybrid method has been proven to give better accuracy in classifying cardiac diseases as compared to traditional deep learning or machine learning methods alone. Nagavelli *et al.* (2022) heart failure affects many individuals, and in the early stages of diagnosis, ECGs play an essential role, so this paper will review what already exists in literature regarding machine learning approaches to predict heart diseases with applications of Naive Bayes, SVM, and XGBoost on how it would enhance the ability of healthcare providers to deliver timely interventions.

(Saboor *et al.*, 2022) heart disease detection is the need of the hour for preventing early-stage heart failure from occurring that saves lives. In this paper, they implemented nine machine learning classifiers on a heart disease dataset and used various metrics to assess performance. Outcome reflects that both standardization of data and tuning of hyperparameter improved the accuracy, though not all the results were in its favor. (Noroozi

et al., 2023) The feature selection methods used when developing machine learning algorithms to forecast heart disease based on the Cleveland dataset are evaluated in this research, with varying degrees of success in improving the prediction model. With an accuracy of 85.5%, SVM-based filtering produced the best results. In general, filtering techniques are superior. The use of wrapper and evolutionary techniques improves specificity and sensitivity.

3. Data Sources

The dataset from the continuing research of the heart is downloaded from the Faisalabad Institute of Cardiology (FIC) website (https://www.fic.gop.pk), strive to be a leader in the provision of Evidence-Based Healthcare to patients with cardiac conditions and in the academic preparation of healthcare providers for the difficulties of the future. There are 304 patient records with different types of information. All attribute listed in table 1.

Table 1: Attribute Used in Dataset

Attribute	Description	Normal Range
Sex	Gender of patients	1= Male 0= Female
Trtbps	Resting Blood Pressure	<120mmHg
Chol	Cholesterol level	<200mg/dL
Fbs	Fasting Blood Sugar level	100-140mg/dL
HTN	hypertension	130-140mmHg
Target	Heart Disease	___

4. Methodology

This study will examine a number of machine learning techniques, including Random Forest, Support Vector Machine, KNN, Logistic Regression, Naive Bayes, and Decision Trees, that are intended to predict cardiac illness. Methodology this section outlines a framework for the proposed model that is used in transforming raw data into recognizable patterns for user understanding of the said process (Jindal *et al.*, 2021). The steps of the process will

include data mining, extraction of significant values, and preprocessing of the data, which includes missing values, cleaning, and normalization as these vary with usage of algorithms. The procedural programming features of Python were used to organize and carry out the many steps, SciPy, NumPy, Matplotlib and Pandas were the libraries and code packages utilized in this study. Several classifiers are used to the preprocessed data, including

KNN, Random Forest, Naive Bayes, Support Vector Machine, Logistic Regression, and Decision Trees. And lastly, some metrics are used in conclusion for accuracy of the model as well as performance. The learnt model is evaluated by matching the symptoms to predict heart disease. We estimate at this point whether or not a patient has heart disease. The model is used if the predicted precision is up to the specified level.

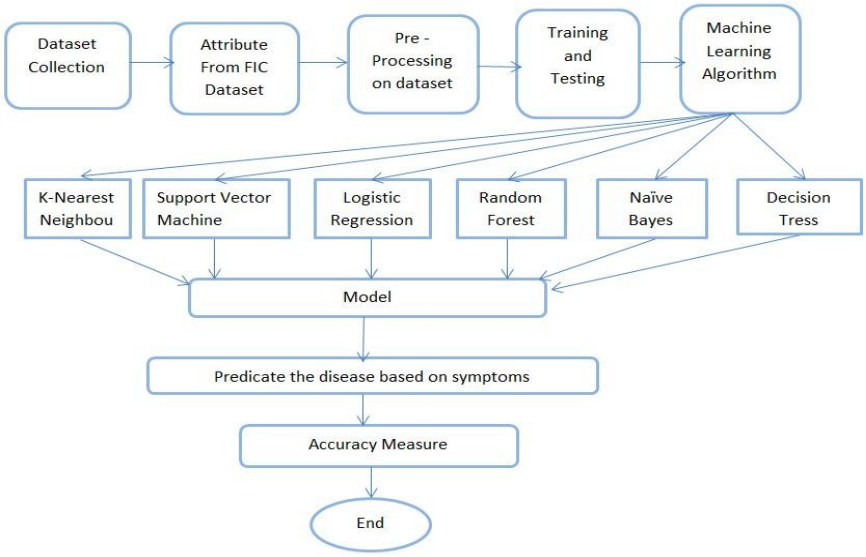

Figure 1. Proposed System

5. Results

This section reports the experimental outcomes acquired by using different supervised learning algorithm. The accuracy of heart disease prediction was determined in the experiment below using an SVM, KNN, LR, NB, and DT. In this paper, we employed RF to present a useful method for assessing cardiac syndrome. For the cardiac data set, the method we suggest RF has an accuracy of

86.89%. RF has increased the accuracy of heart disease prediction by utilizing fewer parameters. In this instance, the Logistic Regression approach with this dataset produced results that were almost as accurate as our intended model. Our approach uses fewer parameters and provides the most accurate classification of heart disease compared to previous techniques.

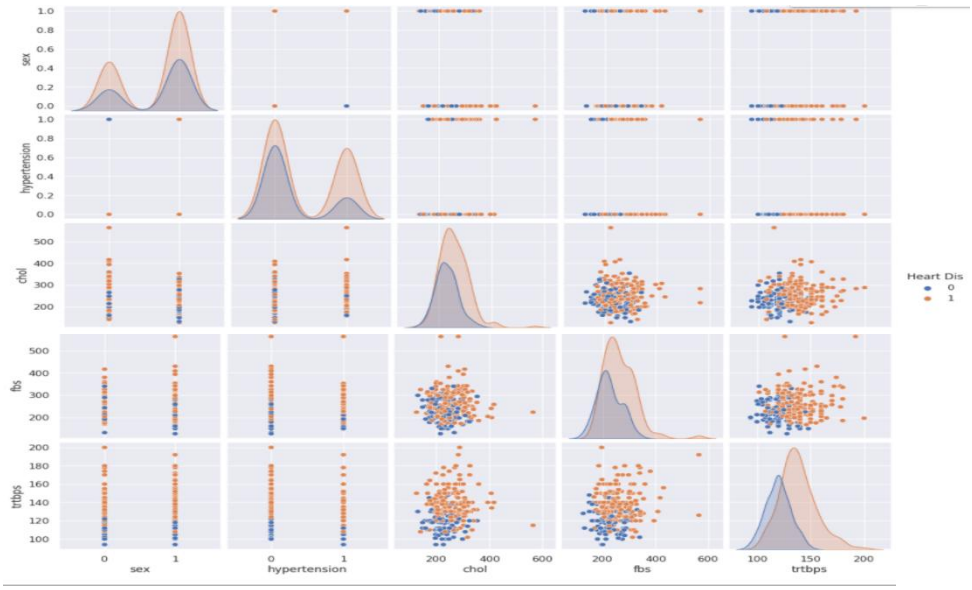

Figure 2: Representation of Dataset with/without Heart Disease

Using "Target" as the hue, the pair plot represents data distribution differences between patients having and not having heart disease as 65.02% and 34.98% respectively. This visualization will bring out different relationships that could be predictors of having heart disease.

Table 2: Accuracy Obtain From the Evaluation Experiments

Classification Algorithm	Accuracy
Random Forest	86.89%
Navie Bayes	78.69%
Support Vector Machine	77.05%
K-Nearet Negibour	83.61%

Decision Tree	75.41%
Logistic Regreesion	85.25%

Table 2 Accuracy of six prediction models compared to our proposed model, in which our proposed model gets the maximum accuracy of 86.89%. Thus, the other techniques using these five attributes get less accuracy compared to our proposed model.

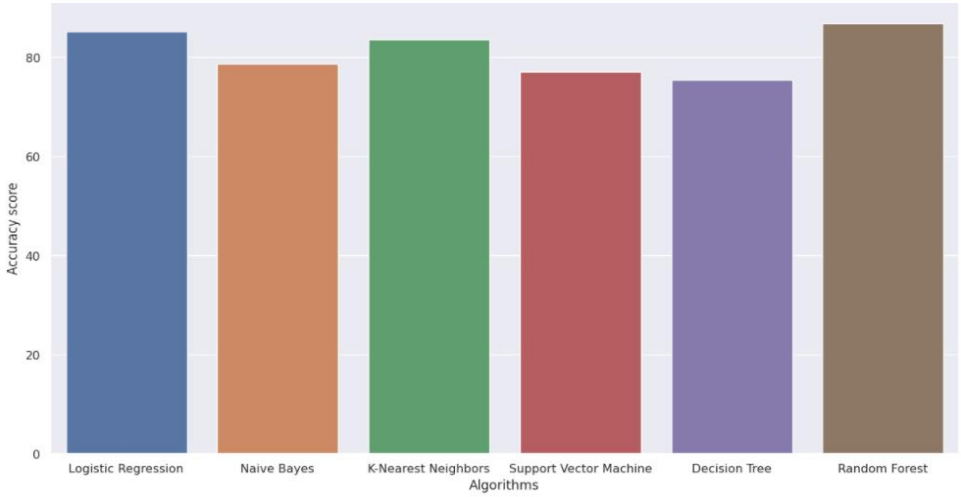

Figure 3: Comparing Accuracy across Different Algorithms

The bar plot indicates accuracy scores for various algorithms. In this experiment, Random Forest achieved the highest accuracy for heart disease prediction. This would indicate that the algorithms have a high difference in their performance and that Random Forest is more accurate than others.

6. Conclusion and Future work

The heart is thought to be the most critical organ in the human body, making the prediction of cardiac disorders one of the most significant health issues. High accuracy algorithms would therefore be a key criterion for assessing their effectiveness in machine learning. The quality of the dataset used for training and testing has a significant impact on this accuracy. Data was first processed, and any extraneous data was discarded. Using the necessary data, all six approaches were applied to forecast heart diseases using only the pertinent attributes. The Google Collaborator tool was

used for this, with accuracy rates ranging from 86.89% for Random Forest, 77.05% for Support Vector Machine, 83.61% for KNN, 85.25% for Logistic Regression, 78.69% for Naive Bayes, and 75.41% for Decision Tree. For the cardiac data set, the method we suggest RF has an accuracy of 86.89%. RF has increased the forecast accuracy of heart disease by utilizing fewer parameters. Even though accuracy increased, much more work needs to be done going forward. More information could be gathered to improve the accuracy.

7. References

Angraal, S., B. Mortazavi, A. Gupta, R. Khera, T. Ahmad, N.R. Desai, D.L. Jacoby, F.A. Masoudi, J.A. Spertus and H.M. Krumholz. 2020. Machine Learning Prediction of Mortality and Hospitalization in Heart Failure with Preserved Ejection Fraction. *JACC. Heart Fail.*

Ansarullah, S.I., S.M. Saif, P. Kumar and M.M. Kirmani. 2022. Significance of Visible Non-Invasive Risk Attributes for the Initial Prediction of Heart Disease Using Different Machine Learning Techniques. *Comput. Intell. Neurosci.* 2022.

Bhavekar, G.S. and A. Das Goswami. 2022. A hybrid model for heart disease prediction using recurrent neural network and long short term memory. *Int. J. Inf. Technol.* 14:1781–1789.

Diwakar, M., A. Tripathi, K. Joshi, M. Memoria, P. Singh and N. Kumar. 2020. Latest trends on heart disease prediction using machine learning and image fusion. *Materials Today: Proceedings*. Elsevier Ltd. pp.3213–3218.

Jiang, S. 2020. *Heart Disease Prediction Using Machine Learning Algorithms*.

Jindal, H., S. Agrawal, R. Khera, R. Jain and P. Nagrath. 2021. Heart disease prediction using machine learning algorithms. *IOP Conference Series: Materials Science and Engineering*. IOP Publishing Ltd.

Katarya, R. and S.K. Meena. 2021. Machine Learning Techniques for

Heart Disease Prediction: A Comparative Study and Analysis. *Health Technol. (Berl)*. 11:87–97.

Kompella, S., V. Boddu and K. Subhadra. n.d. *Neural network based intelligent system for predicting heart disease*.

Mahesh, B. 2020. Machine Learning Algorithms - A Review. *Int. J. Sci. Res.* 9:381–386.

Nagavelli, U., D. Samanta and P. Chakraborty. 2022. Machine Learning Technology-Based Heart Disease Detection Models. *J. Healthc. Eng.* 2022.

Noroozi, Z., A. Orooji and L. Erfannia. 2023. Analyzing the impact of feature selection methods on machine learning algorithms for heart disease prediction. *Sci. Rep.* 13.

P. U. Anitha, BV Pranay Kumar and Ch. Prudvini. 2022. Prediction of Heart Disease using Machine Learning Algorithm: Support Vector Machine. *Int. J. Adv. Res. Sci. Commun. Technol.* 2:166–174.

Pal, M. and S. Parija. 2021. Prediction of Heart Diseases using Random Forest. *Journal of Physics: Conference Series*. IOP Publishing Ltd.

Patro, S.P., G.S. Nayak and N. Padhy. 2021. Heart disease prediction by using novel optimization algorithm: A supervised learning prospective.

Saboor, A., M. Usman, S. Ali, A. Samad, M.F. Abrar and N. Ullah. 2022. A Method for Improving Prediction of Human Heart Disease Using Machine Learning Algorithms. *Mob. Inf. Syst.* 2022.

Saleh Alotaibi, F. 2019. *Implementation of Machine Learning Model to Predict Heart Failure Disease*.

Shah, D., S. Patel and S.K. Bharti. 2020. Heart Disease Prediction using Machine Learning Techniques. *SN Comput. Sci.* 1:345.

Singh, A. and R. Kumar. 2020. Heart Disease Prediction Using Machine Learning Algorithms. *2020 International Conference*

on Electrical and Electronics Engineering (ICE3). pp.452–457.

Sujatha, P. and K. Mahalakshmi. 2020. Performance Evaluation of Supervised Machine Learning Algorithms in Prediction of Heart Disease. *2020 IEEE International Conference for Innovation in Technology (INOCON)*. pp.1–7.

Tarawneh, M. and O. Embarak. 2019. Hybrid Approach for Heart Disease Prediction Using Data Mining Techniques. *Lecture Notes on Data Engineering and Communications Technologies*. Springer Science and Business Media Deutschland GmbH. pp.447–454.

Tougui, I., A. Jilbab and J. El Mhamdi. n.d. Heart disease classification using data mining tools and machine learning techniques. , doi: 10.1007/s12553-020-00438-1/Published.