Article

A web-based heart disease prediction system using machine learning algorithms

Md. Mahbubur Rahman, **Morshedur Rahman Rana**, **Md. Nur-A-Alam**, **Md. Saikat Islam Khan**, **Khandaker Mohammad Mohi Uddin**

Department of Computer Science and Engineering, Dhaka International University, Dhaka-1205, Bangladesh

E-mail: mahbubur.cse@diu.ac, ranamorshedurrahman@gmail.com, munnacse44@gmail.com, saikat.cse@diu.ac, jilanicsejnu@gmail.com

Received 23 February 2022; *Accepted 11 March 2022*; *Published 1 June 2022*

Abstract

Disease diagnosis is the most critical task in the medical diagnosis system. At present, the biggest challenge is to predict heart disease very quickly; for that limitation, the number of dying people is increasing day by day. If a heart disease is diagnosed quickly, we can reduce the death rate indisputably. Thus, this research produces a manual and web-based automatic prediction system that can confer a conceptual report of clear warning of patient's heart condition. The proposed prediction system predicts heart disease using some health parameters. The system uses thirteen health parameters like age, sex, chest pain type, blood pressure, ECG, etc. Eight algorithms are used separately to diagnose heart disease accurately, namely KNN, XgBoost, Logistic Regression (LR), Support Vector Machine (SVM), Ada Boost, Decision tree (DT), Naïve Bayes, and Random Forest (RF). Decision Tree and Random Forest provide better performance than others among all methods. This research also established a website to easily check their heart condition from home instantly. The system has used 1026 individual patients' data for training and testing. It achieves higher accuracy in the different algorithms such as DT (99%), RF (99%), XgBoost (95%), KNN (89%), SVM (85%), LR (85%), Ada Boost (83%) and Naïve Bayes (82%). The experiment result provides a target value of 0 or 1 that refers to the patient's presence or absence of heart disease.

Keywords heart disease prediction; machine learning algorithms; web application; health parameters.

Network Biology
ISSN 2220-8879
URL: http://www.iaees.org/publications/journals/nb/online-version.asp
RSS: http://www.iaees.org/publications/journals/nb/rss.xml
E-mail: networkbiology@iaees.org
Editor-in-Chief: WenJun Zhang
Publisher: International Academy of Ecology and Environmental Sciences

1 Introduction

Cardiovascular Disease (CVD) is one of the leading positions holding diseases in the race of global death in developed countries. The mortality rate due to cardiovascular attack is increasing unusually in high-income regions and posing a significant challenge in the health sector (Faizal et al. 2021; Bantham et al., 2021). According to the Survey of National Surveillance System for Cardiovascular Disease (2021), the age-adjusted prevalence of all categories of heart disease is 7%, where coronary artery disease is higher in males (8.3%)

than females (6.1%) patients (Eremiasova et al., 2020). At the same time, coronary heart disease has been the reason for 16% of global death since 2019 (Yuyun et al., 2020). Unhealthy diet, lack of exercise, overweight, high systolic BP, high cholesterol level, greater fasting glucose level, high body mass index, tobacco-based smoking, and physical inactivity are the primary morbidity of all types of heart disease (Willinger et al., 2021). Angina, pressure, and pain in the chest, sometimes shortened breath are counted as the royal prefix of heart disease. Due to obstacles in blood circulation, sometimes patients feel numbness in the neck, jaw, throat, arm, and leg. Cardiovascular heart disease enhances boundless health and economic burdens in the USA and globally. According to WHO research, 8.9 million deaths worldwide in 2019, and the principal cause of those extensive deaths is cardiovascular heart disease. Coronary heart disease has formed the world's biggest killer in the last 15 years (Builder, 2021; Parambi, et al., 2020; Liou et al., 2020). Machine learning approaches have given an excellent solution in analyzing and predicting millions of clinical data. Those techniques also provide the severity of heart disease based on classification algorithms like Decision Tree (DT), K-Nearest Neighbor Algorithm (KNN), Genetic Algorithm (GA), Random Forest (RA), and Naïve Bayes (NB) Algorithms (Yadav et al., 2018, 2019; Bari Antor et al., 2021; Shrivastava et al., 2021; Varshini et al., 2021; Mostafiz et al., 2021). A precise system can easily predict heart disease, so we can easily hope to diagnose heart early; meanwhile, the system can help the medical sector save more lives by predicting. The current method of diagnosing the heart system is expensive and complex. Every patient wants to be checked up again and again to know the condition of the heart, which takes a lot of time. Several tests like chest X-rays, angiography, ECG, etc., are done to see the state of the heart that is very costly. Knowing the heart condition based on some parameters at home will be more efficient and conducive. Thus, the system has established a website through which people can access the website from home and know the condition of the heart by providing some information. This system works in two ways, firstly through the website, we can see our heart condition, and secondly, we can do it manually from any computer that has the system. Heart specialists built massive records of patients' databases and stored those open for future analysis. The system has analyzed standard datasets such as Kaggle, Cleveland, Hungary, and Switzerland heart disease datasets. The proposed system preprocesses the dataset to complete training and testing procedures. The system has taken the best of 14 attributes as input from 76 features. These 14 attributes play a vital role in determining heart condition. Then the research has applied eight different classification approaches and provides two target values that are 0 or 1. The target value Zero means better shape at heart, and one implies the heart condition is not so good.

A vast number of researchers have discovered different heart risk prediction systems. The main goal of every researcher is to predict heart disease easily and early. Thus, they have provided various heart disease prediction techniques so that the mortality rate can be significantly reduced. The main exploit of this research is:

- The proposed system has used 14 significant parameters for testing. The dataset consists of 1026 different patient data.
- Eight classification techniques are performed to predict heart risk conditions. All of the decision trees work better and provide excellent accuracy.
- The system has applied performance metrics such as accuracy, sensitivity, specificity, precision, F-measure, and classification error to measure the system performance.

This paper is structured like this: Section 2 describes the existing research. The research methods and system views are described in Sections 3 and 4. Section 5 analyzes the results and articulates the implementation methods. Finally, the conclusion of the study and future scopes are mentioned in Section 6.

2 Literature Review

Several researchers have mentioned and implemented various detection techniques of heart disease using

different datasets and machine learning approaches. Most of the researchers used some parameters like blood pressure, cholesterol, obesity, sex, unhealthy diet, lack of exercise, being overweight, high systolic Blood Pressure, high body mass index, high cholesterol level, tobacco smoking, high fasting plasma glucose level, and conclude that a heart disease patient can be attacked for those reasons. Budholiya et al. proposed a system where his team uses XgBoost algorithm along with Random Forest (RF), Extra Tree (ET) with Bayesian optimization, One-Hot (OH) encoding techniques that show the higher accuracy from the XgBoost algorithm, which reached about 91.8% (Budholiya et al., 2020). Yadav et al. used 1025 instances with 14 attributes from the UCI repository by using four tree-based classification algorithms: M5P, random Tree, and Reduced Error Pruning with the Random Forest ensemble method. They reached 99% accuracy on Random Forest (RF). They have also used three features-based algorithms: Pearson Correlation, Recursive Features Elimination, and Lasso Regularization (Yadav et al., 2020). Furthermore, Bergamini et al. proposed a Mapping risk of ischemic heart disease (IHD) using machine learning in a Brazilian state. Their main goal was to create and validate a Heart Health Care Index (HHCI) to predict the risk of IHD based on location and risk factors. They've used Secondary data, geographical information systems (GIS), and ML to validate the Heart Health Care Index (HHCI) and found an RMSE of 0.789 and error proportion close to one (0.867) in Support Vector Machine (SVM) (Bergamini et al., 2020). Samhitha et al. raised a system that used outfit characterization techniques to improve the exactness of frail calculations by consolidating different classifiers. They've also executed the calculation with a restorative dataset (Samhitha et al., 2020). Singh et al. planed a system where his team used algorithms like k-nearest neighbor (KNN), decision tree (DT), linear regression, and support vector machine (SVM) by using UCI Heart disease dataset and found 87% accuracy on KNN (Singh et al., 2020). Balakrishnan et al. proposed a model that his team constructed using Deep Neural Network and $\chi 2$-statistical model. They solve the problem of underfitting and overfitting issues and used DNN and ANN to analyze the efficiency of the model, which accurately predicts the presence or absence of heart disease (Balakrishnan et al., 2021). Regression and Classification techniques are used to mine the data of the Cleveland heart dataset by Kavitha et al (2021). Decision Tree, Random Forest, and Hybrid model fused with DT and RF are applied as a classification method. The Hybrid model showed the highest accuracy pointed at 88.7% in this proposed prediction system. Moreover, Terada et al. used three machine learning classifiers in their medical diagnosis support system (MDSS) for atherosclerosis (Terrada et al., 2020). In their proposed system, UCI and Sani Z-Alizadeh, clinical datasets are used in the training and testing phase. The performance was measured by accuracy, precision, recall, F1-score, and Mathew's correlation coefficient and showed that the 10-folds cross-validation method gained 94% accuracy. Fuad Ali Mohammed Al-Yarimi et al. endeavored an effective feature selection and analyzing technique to get actual accuracy to predict heart disease (Al-Yarimi et al., 2021). The variable size n-gram patterns technique examines the feature selection process of the dermatological dataset. The Naïve Bayes classifier is performed on particular optimal features to get the highest accuracy and maximum sensitivity. On the other hand, Mehmood proposed a method named CardioHelp to predict the presence of heart disease at an early stage (Mehmood et al., 2021). This research paper prepared a dataset and compared it with the state-of-art method to improve accuracy. Using 13 parameters, they have shown the accuracy is best in CNN that is 97%. Rani worked with the neural network method for feasible dataset classification. This research has used 13 parameters in its analysis phase and tried to increase the classification process's efficiency (Rani et al., 2011). In 2013, the research conducted by Taneja et al. (2013) said that diagnosing heart disease patients in a timely is the most challenging task for the medical fraternity for high treatment cost and not affordable for all patients. So, the researcher proposed a system to predict heart disease using different supervised machine learning algorithms. The system shows 95.56% accuracy using the J48 technique, 92.42% accuracy in the SMO technique, and 94.85% accuracy in the

multilayer perception approach. Acharya et al. assessed the problem of predicting and forecasting heart rate variability signals using KNN, Naïve Bayes, and SVM algorithms. Using various classifications in his diagnosis test, he has shown the highest accuracy of 92.02% (Acharya et al., 2015). After that, Aman et al. (2021) established a heart disease prediction system using the WEKA tool and 10-Fold cross-validation. The author worked with SMO, J48, KStar, and Bayes Net classification to predict accuracy (Aman et al., 2021). Based on his research performance, the accuracy of SMO is 89%, 87% accuracy in Bayes Net, and accuracy of Multilayer perceptron, J48, and KStar are 86%, 86%, and 75%, respectively. The accuracy achieved from those algorithms is not satisfactory considered to others. Mohan et al. suggested a unique strategy for identifying key characteristics using classification techniques, which enhance the accuracy of cardiovascular disease forecasting. The proposed method for heart disease using the hybrid random forest with a linear model delivers an improved effectiveness with an overall accuracy of 88.7%, according to research (HRFLM) (Mohan et al., 2019). Furthermore, Riyaz et al. proposed a model that predicts heart disease using SVM, Naïve Bayes, Association rule, KNN, ANN, and Decision Tree classification techniques (Riyaz et al., 2022). Mamun Al et al. suggested a research based on three-classification based on k-nearest neighbor (KNN), decision tree (DT), and random forests (RF) algorithms utilizing a heart disease dataset gathered from Kaggle. With 100 percent accuracy, sensitivity, and specificity, the RF technique was successful (Ali et al., 2021a). Ali and Bukhari provided a model with a 93.33 percent accuracy rate, and that is greater than the accuracy of twenty-eight recently established HF risk prediction models, which ranged from 57.85 percent to 92.31 percent. If this method is used in a clinical setting, this will be beneficial to physician (Ali et al., 2021b).

3 Proposed Methodology

In our proposed method, heart disease can be detected more efficiently and less costly within a short time. This paper worked with preprocessed data to train and test using machine learning algorithms. In the first stage, preprocessed data are divided into two parts. Most of those are used in the training phase (80%), and the rest (20%) are used in the testing phase. In the training and testing phase, the proposed system has trained our dataset using machine learning algorithms like Decision tree, XgBoost, KNN, Support vector machine, Naïve Bayes, Logistic Regression, AdaBoost, and Random Forest. Using the Jupyter platform, we have trained and finally predicted the result of a patient. As shown in Fig. 1, our trained system will indicate the presence of cardiovascular disease in a patient. The primary intention of this raised system is to detect heart disease more efficiently and accurately.

3.1 Preprocessing

Firstly, we have preprocessed the collected dataset to reduce complexity and enhance user accessibility. There are many datasets in Kaggle, UCI, Cleveland, etc. In those entire data centers, more than 76 attributes are presented. We have selected the Kaggle dataset, where 13 parameters are assigned, and those factors are more responsible for heart disease. The selected Kaggle dataset contains almost equally condition and non-disease patient data, whereas the UCI dataset is not similarly divided for disease and non-disease patients. That diminishes the number of inputs to the network and helps it learn more accurately and efficiently. The most appropriate top 13 attributes are age, sex, chest pain, blood pressure, serum cholesterol, fasting blood sugar, resting ECG, thalassemia, max heart rate achieved, ST depression induced by exercise relative to rest, significant vessels and using those factors the system predicts the heart condition. The following Table 1 represents in detail our input attributes.

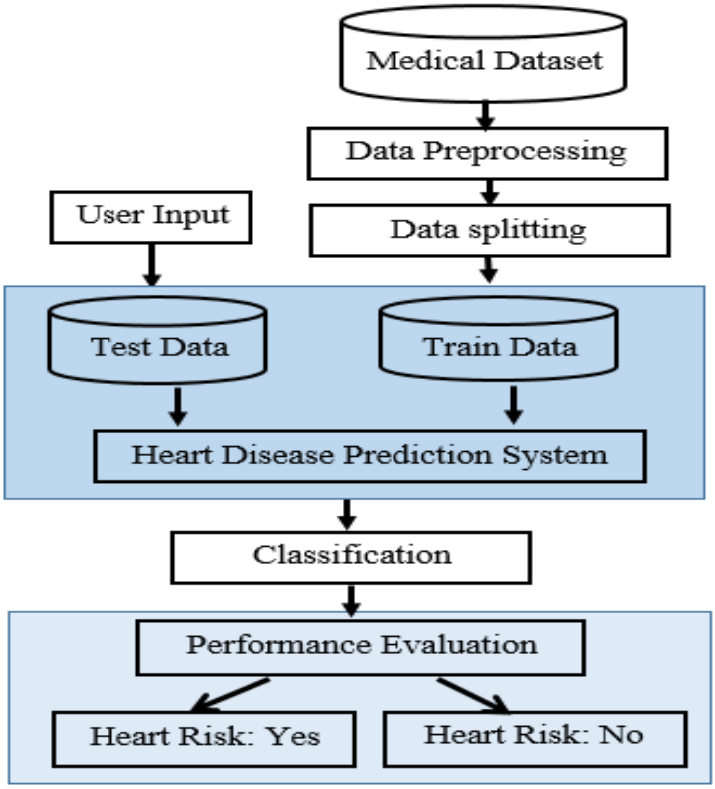

Fig. 1 Proposed Heart Prediction System.

Table 1 Representation of input attributes.

Input Attributes	**Description of Attributes**	**Data Types**	**Format**
Age	Age of	Numerical value	Any floating value
Sex	Gender of patient	Binary value	For male=1, female=0
Chest Pain	4 types of chest pain	Numerical value	For typical angina=1, atypical angina=2, non-angina pain=3, asymptotic=4.
Testbps	Measurement of blood pressure	Continuous value	Any floating value in mm/Hg
Cholesterol	The amount of high- and low-density lipoprotein (HDL and LDL) cholesterol.	Continuous value	Any continuous value in mm/dL
Fasting Blood Sugar	Fasting blood sugar value of an individual with 120mg/dl standard	Binary value	If fasting blood sugar > 120mg/dl then: 1 (true) else: 0 (false)
Rest ECG	Resting electrocardiographic results are 3 types like normal, ST-T wave abnormality and left ventricular hypertrophy.	Numerical value	For normal=0, having ST-T wave abnormality=1, left ventricular hypertrophy=2

Thalach	The max heart rate achieved by an individual.	Binary value	Exercise induced angina: 1 = yes 0 = no
Exang	Exercise induced from angina.	Binary value	Exercise induced angina: 1 = yes and 0 = no
Oldpeak	ST depression induced by exercise that represents the state of rest.	Floating value	Any continuous value
Slope	Slope value during exercise that is measured from ST segment.	Nominal value	Peak exercise ST segment: 1 = up sloping; 2 = flat 3 = down sloping
Ca	Number of major vessels 0 to 3 that is colored by fluoroscopy.	Numerical value	Number of major vessels from 0 to 3.
Thal	Represent the heart rate of patient in three distinct values.	Nominal value	For normal=3, fixed defect=6, reversible defect=7
Result	Predicted outcome from system.	Binary value	Absence of heart disease =0, Present of heart disease=1

3.2 Splitting

The accurate classification result of the dataset depends on the training and testing phase. To get a better result, we divided our whole dataset into two parts: the majority percent of the dataset (80%) for training, and the rest of those are for testing (20%).

3.3 Classification models

The training data uses eight machine-learning algorithms, i.e., DT, LR, Naïve Bayes, AdaBoost, SVM, RF, XgBoost, and KNN. Each algorithm is explained in detail below.

3.3.1 Decision Tree

A decision tree is one kind of supervised learning algorithm. A decision tree is constructed based on high entropy inputs (Charbuty et al., 2021). It uses tree representation starting from root to edge leaf node, and all leaf node corresponds to a class label, and attributes are depicted on the internal node of the tree. In an entropy system, the decision tree removed irrelevant samples from the dataset, and gained information is known as root. The entropy is as follows:

$$\text{Entropy} = -\sum_{k=1}^{n} p_{jk} \log 2\, p_{jk} \quad (1)$$

Here, k is the response variable, and p_{jk} is the ratio of i^{th} class to a total model count.

3.3.2 KNN

In machine learning, KNN is used to classify as a non-parametric method. That means KNN does not make presumptions about the data distribution used in the analysis. This algorithm predicts the class of a new instance based on the most votes by its closest neighbors. It uses Euclidean distance to measure the length of an attribute from its neighbor. The result can vary for the K value and give the best predictions for the optimal K value. The nearest neighbor is found when the starting value of K is 1. Distance functions measure the value of K. The distance function formula is as follows:

$$\text{Distance} = \sqrt{\sum_{i=1}^{k}(xi - yi)^2} \tag{2}$$

3.3.3 Random Forests

Random forests or random decision forests are an ensemble learning method by which multiple decision trees are built for the result that gives a mean prediction of the individual trees. It corrects the decision trees overfitting problem. But the processes of finding the root node and splitting the feature nodes will run randomly, not like a decision tree.

3.3.4 AdaBoost

AdaBoost is a boosting algorithm that helps to combine multiple weak classifiers into a single robust classifier. This method does not follow bootstrapping. However, it will create different decision trees with a single split (one depth), called decision stumps. The number of decision stumps will depend on the number of features in the dataset. The depth will be measured as weight as follows:

$$\text{Weight } (X_i) = \frac{1}{n} \tag{3}$$

where X_i is the i^{th} training occurrence and n is the count of training occurrence.

3.3.5 Logistic Regression

Logistic regression is similar to linear regression because they both have the same goal of estimating the values of the parameter's coefficients. Unlike linear regression, the output prediction is transformed using a nonlinear function called the logistic function. One can estimate coefficients of logistic function by gradient descent. Logistic regression equation:

$$y = e^{\wedge(b0 + b1*x)} / (1 + e^{\wedge(b0 + b1*x)}) \tag{4}$$

Here, y for predicted output, b0 is the bias, and b1 is the coefficient for the single input value (x).

3.3.6 Naive Bayes

Naive Bayes is a statistical classifier. Naïve Bayes established the probability of one occurrence that occurred for another actual event. The Gaussian function trains the model with prior probability and posterior probability. Bayes' theorem is calculated as the following equation:

$$P(c|x) = \frac{P(X|C)P(C)}{P(X)} \tag{5}$$

$$P(c|x)=P(x_1|c)* P(x_2|c)* P(x_3|c)*\ldots\ldots\ldots P(x_n|c)*p(c) \tag{6}$$

Here, P(c|x) is the posterior probability of class (c, target) given predictor (x, attributes), and P(x|c) is the probability of predictor given class. P(c) is the anterior probability of type, and P(x) is the prior probability of predictor.

3.3.7 Support Vector Machines (SVM)

Support vector machines can manipulate multiple continuous and categorical variables. SVM constructs a hyperplane in multidimensional space to separate diverse classes. SVM iteratively generates an optimal hyperplane, used to minimize an error. The main idea of SVM is to find a maximum marginal hyperplane (MMH) that best divides the dataset into classes.

$$f(X) = w^T + b \quad (7)$$

where w is a dimensional coefficient vector and b is an offset. A subsequent optimization problem can solve that.

3.3.8 XgBoost

XgBoost is a method of decision trees designed on gradient boosted for calculating speed and performance. XgBoost provides collateral tree extending that quickly and accurately solves many data mining problems.

4 System View

The proposed working model is a helpful and less time-consuming system than others. The system reduces treatment costs by providing an initial diagnosis in time. Patients can detect their heart condition using our computer-aided system or website. Using the proposed system, anyone can checkup their heart condition daily without the cost of enormous money and time. Section 4.1 explains a detailed explanation of our manual checkup system, and 4.2 explains the proposed system's web-based checking procedures.

4.1 Manual approach

The computer-aided general system is developed with Jupyter IDE of python. This is available in Anaconda navigator. Any patient can know heart condition using a manual diagnosis system. Fig. 2 shows how a patient can see his heart condition from the initial to the final step. A patient has to input his data into the system to check their cardiac situation. System's trained classification models are ready to show output according to provided data. The system will indicate one's heart is infected or not, and at the same time, they also know the accuracy of their result.

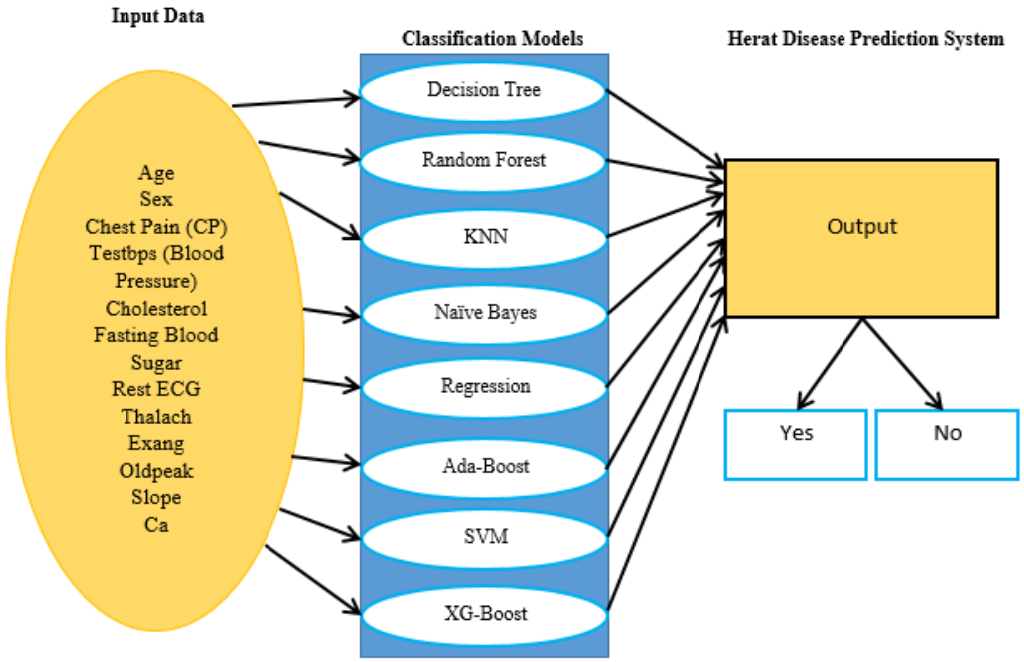

Fig. 2 manual checkup systems to detect heart disease.

The manual checking system is fastened with 13 attributes. Fig. 3 represents the accuracies for given data using eight algorithms. Though all methods will give the same results, accuracy will be different.

4.2 Web based application

The proposed method provides a web-based checking system also for patients from home. Fig. 4 represents the data flow diagram of the heart disease prediction system. To use the system conveniently, the user must log into the system.

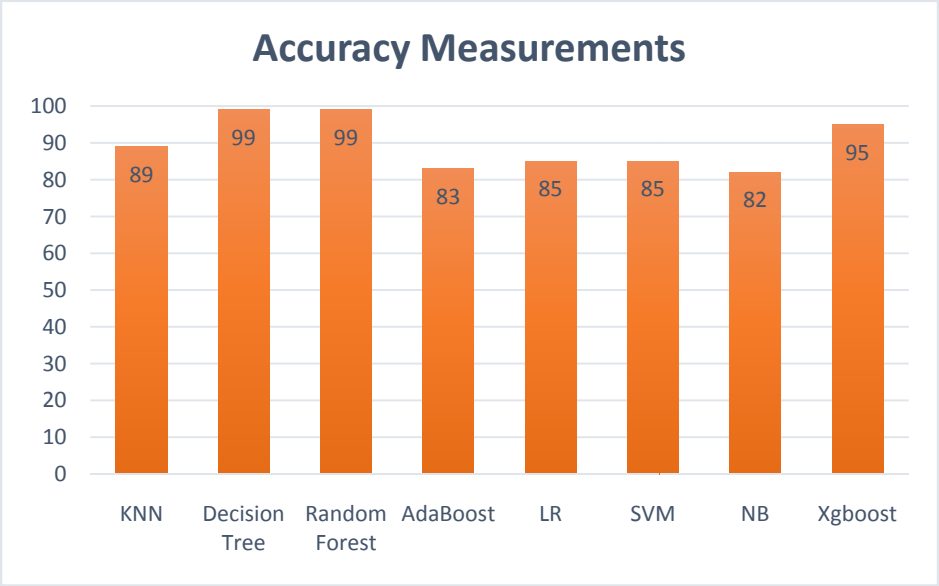

Fig. 3 Manual accuracy measurements.

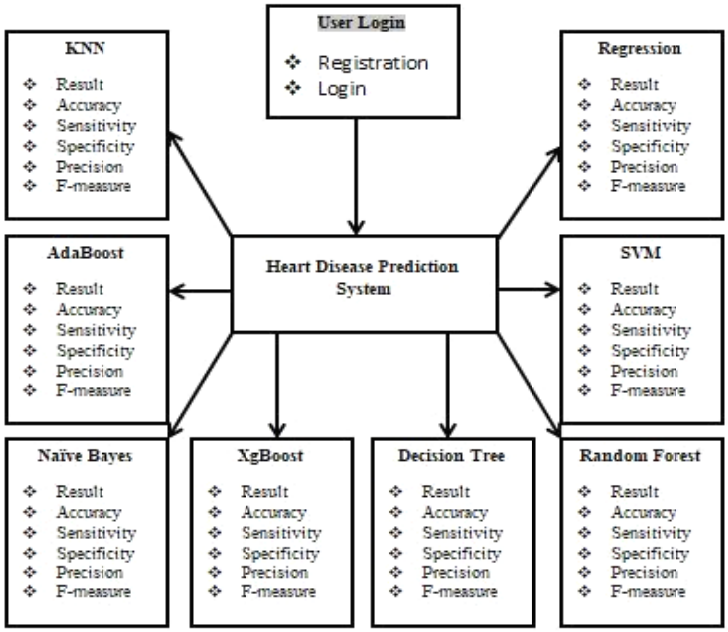

Fig. 4 Data Flow diagram of Heart Disease Prediction System.

The proposed system provides a registration and login form for the users. At first, the patient should create a user id to check their disease. Fig. 5 shows the parameters of the login and registration form. Fig. 6 represents the most relevant 13 parameters of cardiovascular disease. A user will provide some values to the system according to parameters to check their disease. Fig. 7 illustrates the final result of the raised prediction system. The patient can see their heart condition according to machine learning classification methods. Each method shows the accuracy and outcome of that algorithm. Every algorithm will show the exact result but the nearest accuracy.

Fig. 5 User registration and login form.

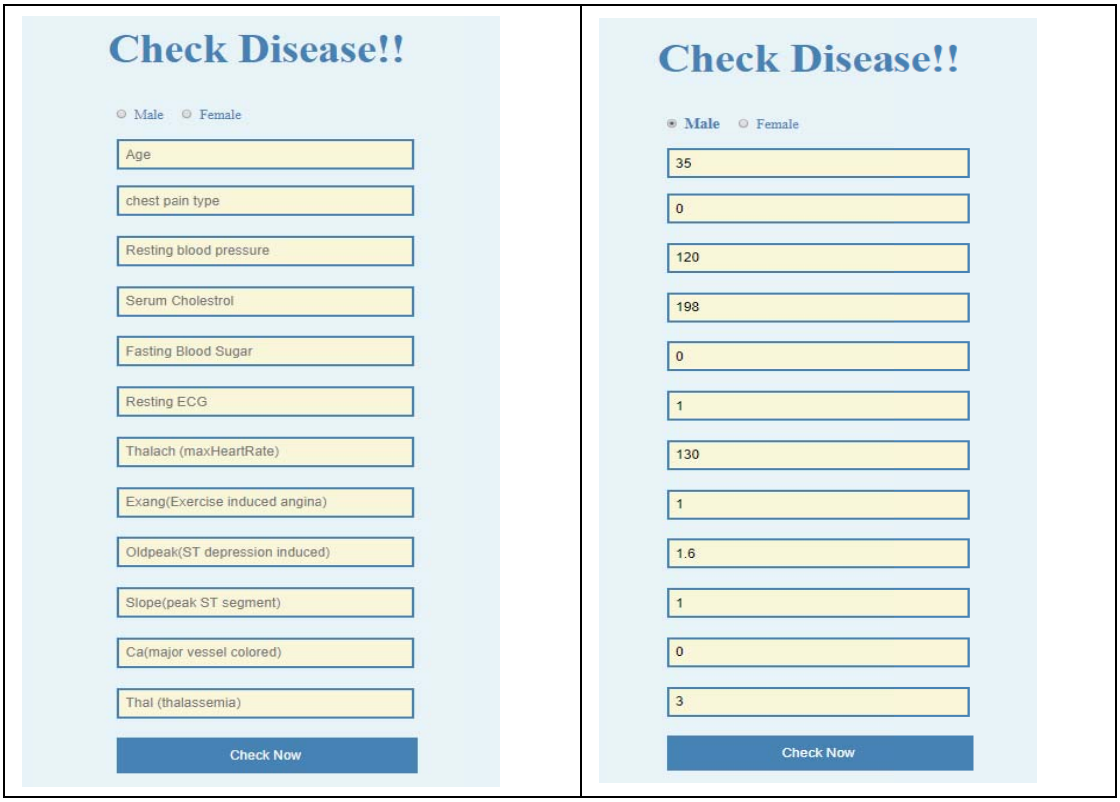

Fig. 6 Attributes of HDPS.

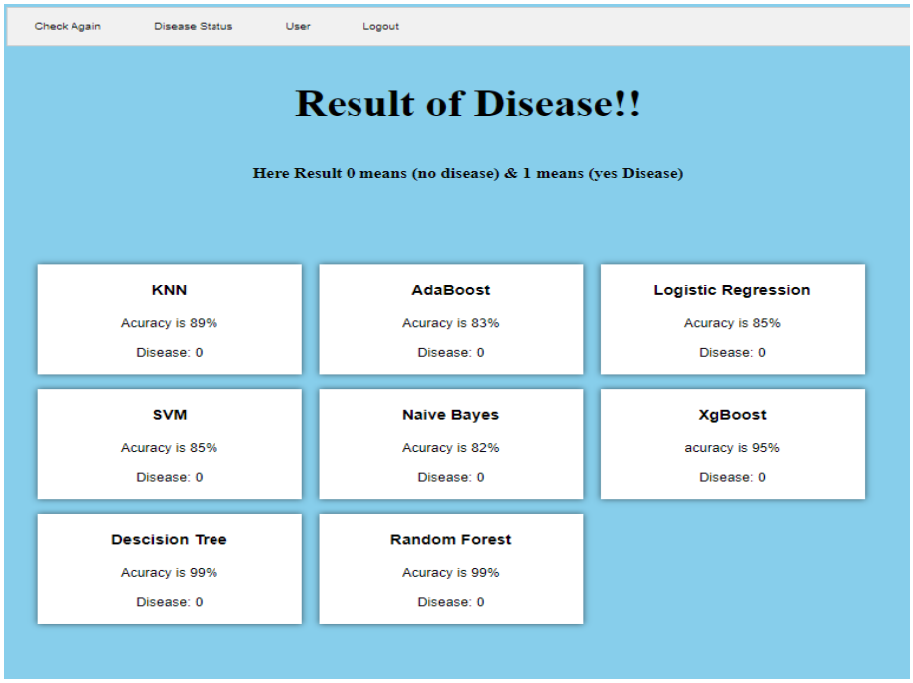

Fig. 7 Results and accuracy of HDPS.

Table 2 Performance matrices equations.

Metrics	***Computing equation***	
Accuracy	$\frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$	**(8)**
Sensitivity	$\frac{TP}{TP + FN}$	**(9)**
Specificity	$\frac{TN}{TN + FP}$	**(10)**
Precision	$\frac{TP}{TP + FP}$	**(11)**
F-measure	$2 * (\frac{Sensitivity * Precision}{Sensitivity + Precision})$	**(12)**
Classification Error	$\frac{\mathrm{FP} + \mathrm{FN}}{FP + FN + TP + TN}$	**(13)**

5 Experimental Results and Analysis

This research has used six measurement schemas like accuracy, sensitivity, specificity, precision, F-measure, and Classification errors to check this system's performance. For calculating six matrices, four several parameters are used. These four parameters are known as True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). Table 2 presents the performance metrics equation. This system has trained on heart disease database and calculated four parameters' values. Then research computed the six measurement schemas and got traditional values representing the proposed system's performance.

Table 3 shows the accuracy, sensitivity, specificity, precision, F-measure, and classification error of different algorithms.

The decision tree has the highest accuracy (99%), sensitivity (98%), and lowest classification errors for the same dataset.

Table 3 Summary of implemented algorithms performance.

Models	Accuracy	Sensitivity	Specificity	Precision	F-measure	Classification Error
KNN	89	90	87	88	88.98	0.112
AdaBoost	83	88	80	79	83.25	0.165
Logistic Regression	85	89	83	80	84.26	0.146
SVM	85	90	82	79	84.14	0.146
Naïve Bayes	82	75	91	90.5	82.02	0.175
XgBoost	95	95	94	94	95	0.053
Random Forest	99	97	96	95	95.5	0.0146
Decision Tree	**99**	**98**	**96**	**97**	**97**	**0.0121**

Fig. 8 illustrates the comparison among eight classification algorithms on accuracy and precision. Among all classifiers, the Decision Tree has given a better performance.

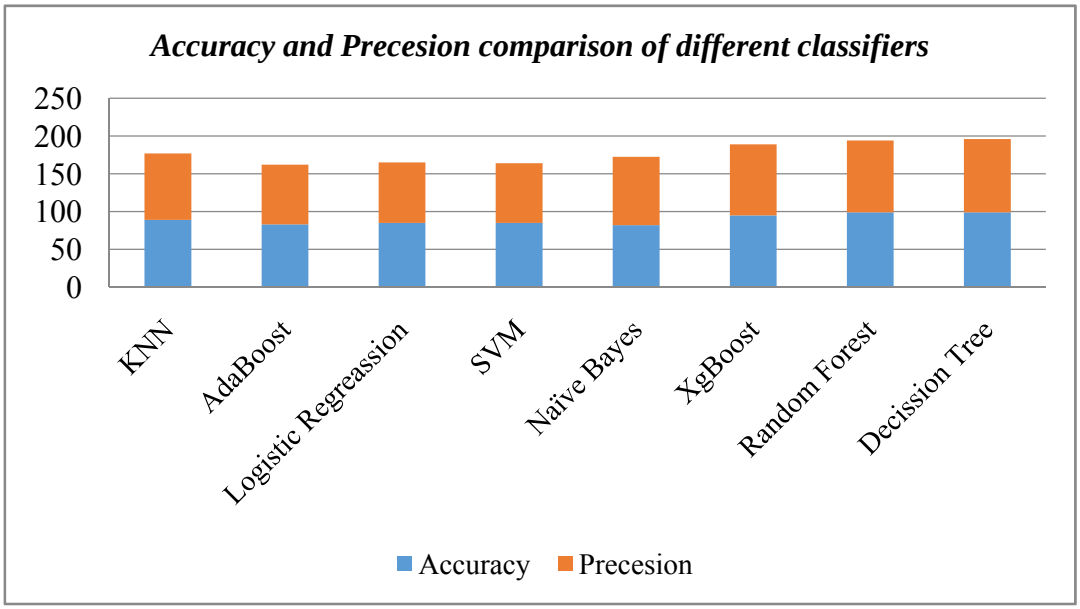

Fig. 8 Comparison of different classifiers.

Table 4 narrates the existing methodologies with their taken attributes and accuracies. From 2006 to now, researchers have been trying to predict heart disease with a computer-aided system to reduce time and cost-efficiently.

Table 4 A comparative studies on various proposed algorithms.

Author	**Proposed Methodology**	**Parameters/ Features**	**Accuracy**
Alim et al. (2020)	Prediction of heart disease was measured with the help of a support vector machine and kernel equivalent to it. ***Techniques used:*** Hoeffding tree method	Not defined	86.94%
Kanwal et al. (2021)	They use Co-Active Neuro-Fuzzy implication method (CANFIS) and later combined with the genetic algorithm to identify heart disease. ***Techniques used:*** Genetic Algorithm,NN, LR,,SVM	14	96%
Khan et al. (2021)	A heart diseases prediction system (HDPS) based on the data mining approaches. ***Techniques used:*** Naïve Bayes, J48 DT, NN, RF.	14	97.70%
Acharya et al.(2015)	A proposed method on heart rate variability signals using data mining techniques where DT provides the highest accuracy. ***Techniques used:*** Naïve Bayes, KNN, DT	8	92.02%
Dwivedi (2018)	Using an artificial neural network and support vector machine, the proposed method predicts stroke patients where SVM provides the highest accuracy. ***Techniques used:*** ANN, SVM.	8	85%
Ayon et al. (2020)	Logistic Regression (LR), Support Vector Machine (SVM), Deep Neural Network (DNN), Decision Tree (DT), Nave Bayes (NB), Random Forest (RF), and K-Nearest Neighbor (K-NN) were the seven computational intelligence techniques used. ***Techniques used:*** DT, RF, NN, Logistic Regression, SVM, Naïve Bayes	9	98.15%
Beyene et al. (2018)	Propose methodology evaluate performance using tenfold cross-validation to predict heart disease. In this method, logistic regression provides the highest accuracy. ***Techniques used:*** Naïve Bayes, SVM, Classification Tree, Logistic Regression, KNN, ANN	12	Not defined
Katarya et al. (2020)	Chronic prediction system using data mining techniques where Naïve Bayes, Decision Tree provide the highest accuracy. ***Techniques used:*** Naïve Bayes, DT, SVM	10	95.56%
Motarwar et al. (2020)	Cardiovascular disease prediction using data mining techniques. Simple Cart provides the highest accuracy. ***Techniques used:*** Naïve Bayes, J48, Simple CART	Not Defined	92.2%

Farzana et al. (2020)	The proposed model predicts heart disease using classification techniques where the SVM technique is more effective and efficient than other data mining algorithms. ***Techniques used:*** Naïve Bayes, KNN Associate Rule, , SVM, ANN, DT	13	95%
Ismail et al. (2020)	A conceptual method to enhance prediction of heart disease using big data where SVM provides the highest accuracy. ***Techniques used:*** Naïve Bayes, SVM	13	90.6%
Sharma et al. (2020)	Data mining techniques predict heart diseases. ***Techniques used:*** Naïve Bayes, J48, SVM	13	90.78%
Proposed method	Heart disease prediction system from web and manual using machine learning classification. The highest accuracy is obtained from the Decision Tree and Random Forest**.** ***Techniques used:*** DT, RF, Naïve Bayes, KNN, AdaBoost, SVM, Logistic Regression, XgBoost	**13**	**99%**

6 Conclusion and Future Work

The research paper shows the overviews of existing methodologies and literature review of heart disease prediction systems which helps us to improve our method. In our approach, using the heart patients dataset from Alim et al. (2020), we analyzed different machine learning classification algorithms to predict the heart disease of any patient manually and on the web platform. The analysis shows 99% accuracy in Decision Tree and Random Forest techniques than other algorithms. Decision Tree (0.0121) has a minor classification error between these methods than Random Forest (0.0146). Further extension of this work is to get 100% accuracy to detect heart disease using more updated machine learning techniques. For enhancing user accessibility, research will be extended by developing an android app.

Abbreviations

BP, Blood Pressure; PA, Physical Exercise; RF, Random Forest; LR, Logistic Regression; HDPS, Heart Disease Prediction System; ANN, Artificial Neural Network; FP, false positive; SVM, support vector machine; TP, true positive; DT, Decision Tree.

Acknowledgement

We are grateful to Professor Dr. H. I. Lutfor Rahman Khan, Head of Cardiology, Dhaka Medical College and Hospital, for his unswerving loyalty and continuous support.

References

Acharya UR, Vidya KS, Ghista DN, Lim WJE, Molinari F, Sankaranarayanan M. 2015. Computer-aided diagnosis of diabetic subjects by heart rate variability signals using discrete wavelet transform method. Knowledge-Based Systems, 81: 56-64

Ali MM, Paul BK, Ahmed K, Bui FM, Quinn JM, Moni MA. 2021a. Heart disease prediction using supervised machine learning algorithms: performance analysis and comparison. Computers in Biology and Medicine, 136: 104672

Ali L, Bukhari SAC. 2021b. An approach based on mutually informed neural networks to optimize the generalization capabilities of decision support systems developed for heart failure prediction. Irbm, 42(5): 345-352

Alim MA, Habib S, Farooq Y, Rafay A. 2020. Robust Heart Disease Prediction: A Novel Approach Based On Significant Feature and Ensemble Learning Model. In: 2020 3rd International Conference on Computing, Mathematics and Engineering Technologies (ICoMET). IEEE

Al-Yarimi FAM, Munassar NMA, Bamashmos MHM, Ali MYS. 2021. Feature optimization by discrete weights for heart disease prediction using supervised learning. Soft Computing, 25(3): 1821-1831

Ayon SI, Islam MM, Hossain MR. 2020. Coronary artery heart disease prediction: a comparative study of computational intelligence techniques. IETE Journal of Research, 1-20

Aman RSC. 2021. Analyzing predictive algorithms in data mining for cardiovascular disease using WEKA tool. International Journal of Advanced Computer Science and Applications, 12(8)

Balakrishnan M, Christopher AA, Ramprakash P, Logeswari A., 2021. February. prediction of cardiovascular disease using machine learning. Journal of Physics: Conference Series, 1767(1): 012013

Bantham A, Ross SET, Sebastião E, Hall G. 2021. Overcoming barriers to physical activity in underserved populations. Progress in Cardiovascular Diseases, 64: 64-71

Bari Antor M, Jamil AHM, Mamtaz M, Monirujjaman Khan M, Aljahdali,S, Kaur M, Singh P, Masud M. 2021. A comparative analysis of machine learning algorithms to predict Alzheimer's disease. Journal of Healthcare Engineering, 2021

Bergamini M, Iora PH, Rocha TAH, Tchuisseu YP, Dutra ADC, Scheidt JFHC, Nihei OK, de Barros Carvalho, MD, Staton CA, Vissoci JRN, de Andrade L. 2020. Mapping risk of ischemic heart disease using machine learning in a Brazilian state. Plos one, 15(12): e0243558

Beyene C, Kamat P. 2018. Survey on prediction and analysis the occurrence of heart disease using data mining techniques. International Journal of Pure and Applied Mathematics, 118(8): 165-174

Budholiya K, Shrivastava SK, Sharma V. 2020. An optimized XGBoost based diagnostic system for effective prediction of heart disease. Journal of King Saud University-Computer and Information Sciences. https://doi.org/10.1016/j.jksuci.2020.10.013

Builder V. 2021. Cardiovascular Pathologies and Disorders. Mosby's Pathology for Massage Professionals E-Book, Elsevier

Charbuty B, Abdulazeez A. 2021. Classification based on decision tree algorithm for machine learning. Journal of Applied Science and Technology Trends, 2(1): 20-28

Dwivedi AK. 2018. Performance evaluation of different machine learning techniques for prediction of heart disease. Neural Computing and Applications, 29(10): 685-693

Eremiasova L, Hubacek JA, Danzig V, Adamkova V, Mrazova L, Pitha J, Lanska V, Cífková R, Vitek L. 2020. Serum bilirubin in the Czech population—Relationship to the risk of myocardial infarction in males. Circulation Journal, 84(10): 1779-1785

Faizal ASM, Thevarajah TM, Khor SM, Chang SW. 2021. A review of risk prediction models in

cardiovascular disease: conventional approach vs. artificial intelligent approach. Computer Methods and Programs in Biomedicine, 207: 106190

Farzana S, Veeraiah D. 2020, October. Dynamic Heart Disease Prediction using Multi-Machine Learning Techniques. In: 2020 5th International Conference on Computing, Communication and Security (ICCCS). IEEE

Ismail A, Abdlerazek S, El-Henawy IM. 2020. Big data analytics in heart diseases prediction. Journal of Theoretical and Applied Information Technology, 98(11): 15-19

Kanwal S, Rashid J, Nisar MW, Kim J, Hussain A. 2021. July. An Effective Classification Algorithm for Heart Disease Prediction with Genetic Algorithm for Feature Selection. In: 2021 Mohammad Ali Jinnah University International Conference on Computing (MAJICC). IEEE

Kavitha M, Gnaneswar G, Dinesh R, Sai YR, Suraj RS. 2021, Heart disease prediction using hybrid machine learning model. In: 2021 6th International Conference on Inventive Computation Technologies (ICICT). 1329-1333, IEEE

Katarya R, Srinivas P. 2020. Predicting Heart Disease at Early Stages Using Machine Learning: A Survey. In: 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC). 302-305, IEEE

Khan M, Rahman A, Rahman M, Salehin JU, Islam M, Rabbi M. 2021. Efficient data mining techniques for heart disease prediction and comparative analysis of classification algorithms. Asian Journal of Research in Computer Science, 2021: 239056682

Liou L, Kaptoge S. 2020. Association of small, dense LDL-cholesterol concentration and lipoprotein particle characteristics with coronary heart disease: A systematic review and meta-analysis. Plos One, 15(11): e0241993

Mehmood A, Iqbal M, Mehmood Z, Irtaza A, Nawaz M, Nazir T, Masood M. 2021. Prediction of heart disease using deep convolutional neural networks. Arabian Journal for Science and Engineering, 46(4): 3409-3422

Motarwar P, Duraphe A, Suganya G, Premalatha M. 2020. Cognitive Approach for Heart Disease Prediction Using Machine Learning. In: 2020 International Conference on Emerging Trends in Information Technology and Engineering (ICETITE). IEEE

Mostafiz R, Uddin KMM, Uddin MS, et al. 2021. Diagnosis of diabetes: A machine learning paradigm using optimized features. Network Biology, 11(3): 222-240

Mohan S, Thirumalai C, Srivastava G. 2019. Effective heart disease prediction using hybrid machine learning techniques. IEEE Access, 7: 81542-81554

Parambi DGT, Unnikrishnan MK, Marathakam A, Mathew B. 2020. Demographic and Epidemiological Aspects of Aging. In: Nutrients and Nutraceuticals for Active & Healthy Ageing. Springer, Singapore

Rani KU. 2011. Analysis of heart diseases dataset using neural network approach. arXiv preprint arXiv:1110.2626

Riyaz L, Butt MA, Zaman M, Ayob O. 2022. Heart Disease Prediction Using Machine Learning Techniques: A Quantitative Review. In: International Conference on Innovative Computing and Communications. 81-94, Springer, Singapore

Samhitha BK, Priya MS, Sanjana C, Mana SC, Jose,J. 2020. Improving The Accuracy in Prediction of Heart Disease Using Machine Learning Algorithms. In: 2020 International Conference on Communication and Signal Processing (ICCSP). 1326-1330, IEEE

Shrivastava K, Jotwani V. 2021. A Comparative Analysis of Various Data Mining Techniques to Predict Heart Disease. Expert Clouds and Applications: Proceedings of ICOECA 2021. Springer

Sharma S, Parmar M. 2020. Heart diseases prediction using deep learning neural network model. International

Journal of Innovative Technology and Exploring Engineering, 9(3): 2244-2248

Singh A, Kumar R. 2020, February. Heart disease prediction using machine learning algorithms. In: 2020 International Conference on Electrical and Electronics Engineering (ICE3). 452-457, IEEE

Taneja A. 2013. Heart disease prediction system using data mining techniques. Oriental Journal of Computer Science and Technology, 6(4): 457-466

Terrada O, Hamida S, Cherradi B, Raihani A, Bouattane O. 2020. Supervised machine learning based medical diagnosis support system for prediction of patients with heart disease. Advances in Science, Technology and Engineering Systems Journal, 5(5): 269-277

Varshini KS, Uthra RA. 2021. An effectual method for disease identification in pediatric dataset. Materials Today: Proceedings. http://dx.doi.org/10.1016/j.matpr.2021.03.317

Willinger L, Brudy L, Meyer M, Oberhoffer-Fritz R, Ewert P, Müller J. 2021. Prognostic value of non-acute high sensitive troponin-T for cardiovascular morbidity and mortality in adults with congenital heart disease: A systematic review. Journal of Cardiology, 78(3): 206-212

Yadav AS, Sharm P, Swami A, Pandey G. 2018. A supply chain management of chemical industry for deteriorating items with warehouse using genetic algorithm. Selforganizology, 5(1-2): 1-9

Yadav AS, Swami A, Kher G. 2019. Blood bank supply chain inventory model for blood collection sites and hospital using genetic algorithm. Selforganizology, 6(3-4): 13-23

Yadav DC, Pal S. 2020. Prediction of heart disease using feature selection and random forest ensemble method. International Journal of Pharmaceutical Research, 12(4): 56-66

Yuyun MF, Sliwa K, Kengne AP, Mocumbi AO, Bukhman G. 2020. Cardiovascular diseases in sub-Saharan Africa compared to high-income countries: an epidemiological perspective. Global Heart, 15(1): 1-18