INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

A Hybrid Machine Learning Model to Predict Heart Disease Accurately

Suresh Subramanian[1]*, **Y Angeline Christobel**[2]*

1 Chairperson, Department of Multimedia, College of IT, Ahlia University, Kingdom of Bahrain
2 Dean, School of Computational Studies, Hindustan College of Arts & Science, Chennai

OPEN ACCESS

Received: 13.01.2022
Accepted: 19.02.2022
Published: 25.03.2022

Citation: Subramanian S, Christobel YA (2022) A Hybrid Machine Learning Model to Predict Heart Disease Accurately. Indian Journal of Science and Technology 15(12): 527-534. https://doi.org/10.17485/IJST/v15i12.104

* **Corresponding authors**.

ssubramanian@ahlia.edu.bh

angelinechristobel5@gmail.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objective: To propose the most effective machine learning algorithm for predicting cardiac problems. **Methods:** The dataset used for this study is "heart" which was taken from www.kaggle.com. The heart dataset contains 13 features and a target variable. It is divided into 70 percent training set and 30 percent testing set. K-Fold cross-validation is used in this study for model evaluation and model selection. The K value chosen is ten. A Hybrid Ensemble machine learning model is built using a heterogeneous collection of weak learners in this work. To construct a hybrid ensemble model, weak learners such as "Logistic Regression", "Decision Tree", "Support Vector Machine", "K-Nearest Neighbor", and "Naive Bayes" are used. Normally, in an ensemble model, a homogeneous group of weak learners is utilized, however in this study, a heterogeneous group of weak learners is used. The parameter used in this study is accuracy. Accuracy of all the weak learners is found and compared with the hybrid ensemble model. **Findings:** Weak machine learning models are combined to create an ensemble model. The "Hybrid Ensemble model" has a 98 percent accuracy rate and outperforms all weak learners such as "Logistic Regression", "Decision Tree", "Support Vector Machine", "K-Nearest Neighbor", and "Naive Bayes". **Novelty and applications :** For the prediction of heart problems, the hybrid ensemble model is recommended since it extracts more accurate and valuable data from huge amounts of data, making prediction easier for physicians.

Keywords: Machine Learning; Ensemble Model; Weak Learners; Homogeneous; Heterogeneous

1 Introduction

Over the last ten years, heart disease has become the top cause of death worldwide. Heart disease is associated with a variety of symptoms, making it challenging to identify it quickly and accurately. Large volumes of healthcare data are collected by the healthcare industries, which must be mined to uncover hidden information for successful decision-making. In the healthcare industry, data science plays a critical role in analyzing vast amounts of data. Machine learning is a type of data analysis that allows computers to learn from data, recognize patterns, and make judgments without the need for human interaction. Heart disease is predicted using a variety of machine learning

methods. For forecasting cardiac disease, some algorithms are quite accurate. As a result, a comparison examination of machine learning algorithms is required to determine which is the most effective in predicting cardiac disease. It aids doctors in detecting cardiac illness at an early stage.

Five categories of machine learning models are used in this research. The models used are "Logistic Regression Model", "Decision Tree", "Support Vector Machine", "K-Nearest Neighbor", and "Naive Bayes". The accuracy of each model is determined and compared to the Hybrid Ensemble model, which is a composite of all five models. The term "Hybrid" is employed because ensemble models use a homogeneous set of machine learning models, whereas this study uses a heterogeneous set of machine learning models.

Many studies have been conducted to predict cardiac disease at an early stage.

Chu-Hsing Lin et al. compared "Convolutional Neural Networks" to "Conventional Neural Networks" to predict heart disease in [(1)]. Their findings demonstrate that the CNN model outperforms the NN model by a factor of 93 percent.

The synthetic minority over-sampling technique was employed by [(2)]. According to their findings, the Fuzzy Random Forest model is the most effective.

Random Forest predicts heart disease more accurately, according to Riddi Kasabe et al. [(3)]. Before being evaluated, the machine learning algorithms were pre-processed.

The importance of care for patients at an early stage was described by Montu Saw et al. [(4)]. According to their findings, the logistic regression model has an accuracy of 80%.

Noor Basha et al. [(5)] examined various machine learning models to predict cardiac illness and discovered that KNN is the most accurate, with an accuracy of 85 percent.

Rahul Kataria et al. [(6)] investigated which feature should be taken into consideration to achieve a better result. For comparison, they employed ANN, Decision Tree, Random Forest, SVM, Nave Bayes, and KNN, and concluded that Random Forest is the best.

Using the SVM and K-NN machine learning methods, heart disease prediction using the risk analysis model was developed by Latin Miao et al. [(7)].

The support vector machine model had a high accuracy than KNN. With 84.28 percent accuracy, Halima El Hamdaoui et al. [(8)] discovered that Nave Bayes is better.

Daniel Ananey-Obiri et al. [(9)] compared three machine learning models, "Linear Regression(LG)", "Decision Tree Classifier (CART)" and "Gaussian Naïve Bayes (GNB)" and found Linear Regression gives good result.

In a study of several machine learning algorithms, Rohit Bharti et al. [(10)] identified Deep Learning, which had an accuracy of 94.28 percent and outperformed all other algorithms.

Muktevi Srivenkatesh employed "K-Nearest Neighbor", "Support Vector Machines", "Logistic Regression", "Naive Bayes", and "Random Forest" in his paper [(11)]. Random Forest, he anticipated, would forecast people with a continuous cardiovascular disappointment infection.

In [(12)], Muhammad Zeeshan Younas compared "Decision Tree", "Logistic Regression", "SVM", "KNN", "Naive Bayes", and "Random Forest", and found that the "Logistic Regression" algorithm was the most accurate, with an accuracy of 86.89 percent.

Abdulwahab Ali Almazroi et al. [(13)] discovered that "decision tree" is the highest performing method when compared to "logistic regression," "support vector machines," and "artificial neural networks." The decision tree is 82 percent accurate.

In [(14)], the authors proposed a new "Multi-Layer Perceptron" for Enhanced Brownian Motion based on Dragonfly Algorithm (MLP-EBMDA) for heart disease classification, as well as an effective unsupervised feature selection technique. In terms of accuracy, the proposed system was compared to various current systems and based on the selected features, the accuracy of the proposed system was found to be 94.28 percent.

Abdulaziz Albahr et al. [(15)] suggested a new computer model for predicting early cardiac disease. The predictive model is supplemented with a new regularisation that decays the weights based on the standard deviation of the weight matrices and compares the outcomes to their parents (RSD-ANN). By a large amount, RSD-ANN beats previous techniques. According to our tests, the average validation accuracy obtained using either the 10-fold cross-validation or holdout technique was 96.30 percent.

2 Methodology

Figure 1 depicts the suggested architecture for this study. Several heterogeneous machine learning models are grouped as weak learners in the proposed method. Individually, weak learners reveal their results across the whole training set, and the results of weak learners are aggregated to provide the final result.

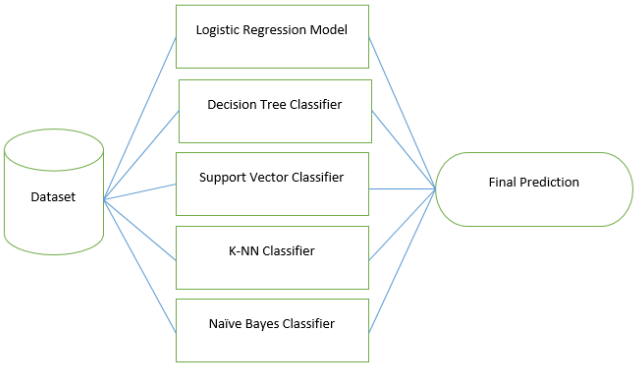

Fig 1. Proposed architecture

2.1 Dataset and Data Interpretation

The dataset used in this study is "heart" which was taken from Kaggle [www.kaggle.com]. The heart dataset contains 13 features and a target variable. The dataset description is given in Table 1.

Table 1. Heart Dataset Description

S.No.	Attributes	Description
1	Age	Age of the patient
2	Sex	Female = "0" and male ="1"
3	Chest Pain(cp)	Chest Pain is categorized into 4 types 0 = " typical angina" 1 = " atypical angina" 2 = " nan-anginal pain" 3 = " asymptotic"
4	Resting Blood Pressure (trestbps)	The patient's resting blood pressure is measured in millimetres of mercury (mmHg) (unit)
5	Cholesterol (chol)	The cholesterol of the patient is in mg/dl (unit).
6	Fasting Blood Sugar (FBS)	Fasting blood sugar is represented as a number between 0 and 1, such as 1 = if fbs >120 mg/dl (true) and 0 = if fbs >120 mg/dl (false) (false).
7	Resting ECG (restecg)	Resting ECG is divided into three types from 0 to 2 defining: 0 = "normal", 1 = "having ST-T wave abnormality", 2 = "left ventricular hypertrophy"
8	Max Heart Rate (thalach)	Maximum heart rate of the patient
9	Exercise-induced angina (exang)	Exercise-induced angina is represented as 0 or 1 such as 0 = "No" and 1 = "Yes"
10	Oldpeak	The value of ST depression is displayed.
11	Slope	The peak of exercise during the ST segment 0 = "up-slope", 1 = "flat", 2 = "down-slope"
12	No. of major vessels (ca)	Colored fluoroscopy is used to classify it into four categories ranging from 0 to 4.
13	Thalassemia (thal)	ranges from 1 to 3, where 1 = "normal", 2 = "fixed defect", 3 = "reversible defect"
14	Target	Prediction attribute 0 = no possibility of heart attack 1= possibility of a heart attack.

Figure 2 shows the pairwise correlation of all columns. The groupings of strongly associated features can be found using pairwise correlation, giving the model more predictive potential.

In Figure 3, the results of an ECG taken when at rest are shown. It is divided into three categories 0: "Normal", 1: "Abnormality in ST-T wave", 2: "Left ventricular hypertrophy" (Nominal)

The highest heart rate of the patients is depicted in Figure 4.

The heart dataset was pre-processed before applying the machine learning models. The standard scalar method is used to pre-process the dataset. The dataset is split into two parts: a training set (70%) and a testing set (30%). This work uses K-fold cross-validation for model evaluation and model selection. The K value chosen is ten. The five machine learning models studied

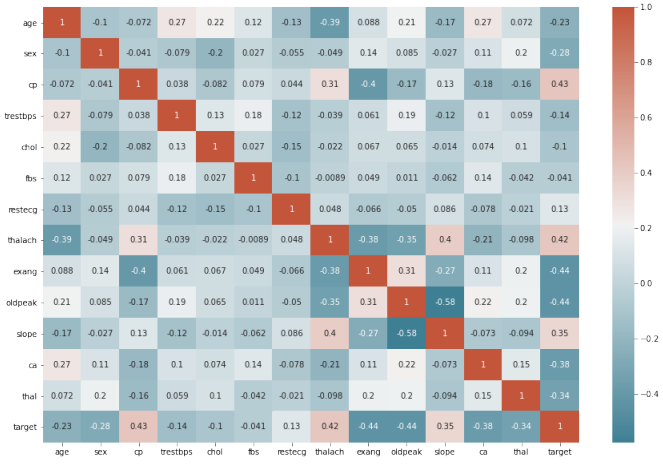

Fig 2. Heat map - pairwise correlation of all columns

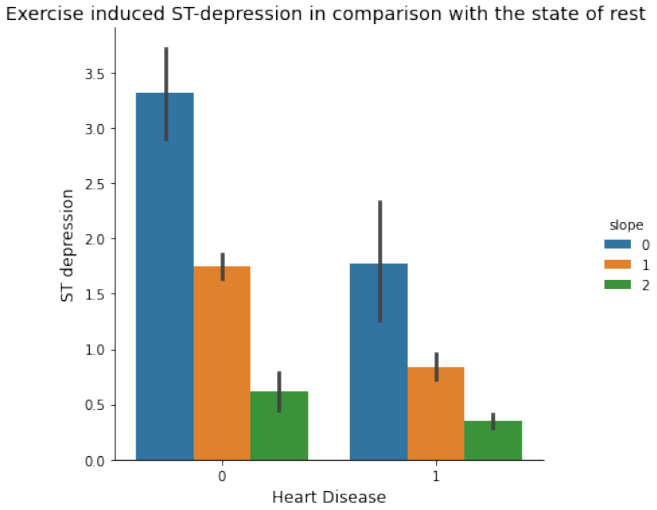

Fig 3. Comparison of ST depression with the state of rest

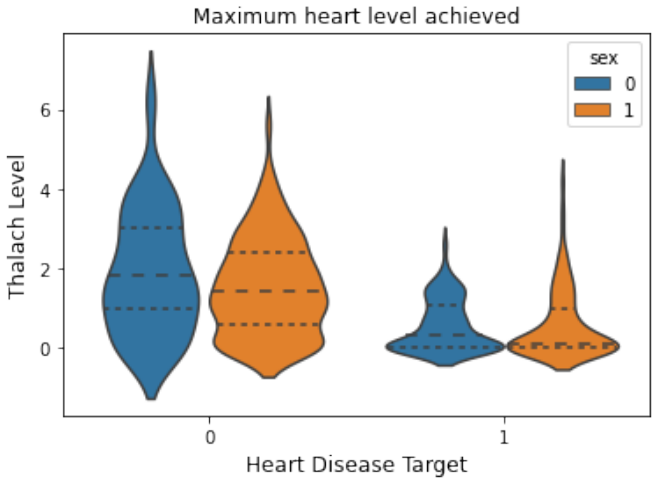

Fig 4. Maximum heart level

in this study are the "Logistic Regression Model", "Decision Tree", "Support Vector Machine", "K-Nearest Neighbor", and "Naive Bayes", and then the "Hybrid Ensemble model" is created by combining these five techniques. A confusion matrix is created as a result of the machine learning models' output, which contrasts the actual target values with those predicted by the machine learning model. The "accuracy" metric is used in this paper to compare the models. For classification accuracy, divide the total number of true predictions by the total number of predictions made. The confusion matrix and formula for calculating accuracy are shown below in Figure 5.

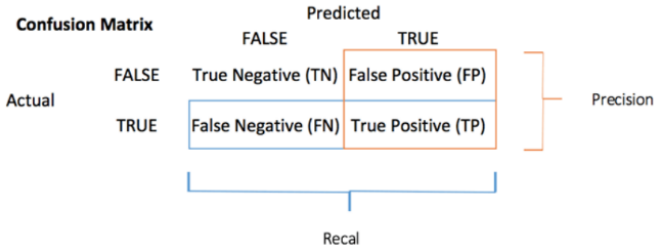

Fig 5. Confusion matrix

2.1.1 K-Fold Cross-Validation

Cross-validation is a technique used in applied machine learning to estimate a machine learning model's skill on unknown data. The process includes only one parameter, k, which specifies the number of groups into which a given data sample should be divided. As a result, the process is frequently referred to as K-fold cross-validation. When a precise value for K is specified, it can be substituted for K in the model's reference, for example, K=10 for 10-fold cross-validation.

The general technique for K-fold validation is as follows:

1. Shuffle the dataset at random
2. Sort the information into K groups
3. Write the following for each separate group

- Use the group as a holdout or test data set.
- Use the remaining groupings as a training data set.
- Create a model for the training set and compare it to the test set.
- Keep the evaluation result but discard the model

4. Summarize the model's ability using the sample of model evaluation ratings

Importantly, each observation in the data sample is assigned to a separate group and stays in that group throughout the technique. This means that each sample has a chance to appear in the holdout set and train the model K several times.

2.1.2 Hybrid Ensemble Model

Bagging is a type of ensemble machine learning strategy that improves performance by combining the outputs of multiple learners. These methods work by dividing the training set into subsets and putting them through several machine-learning models, then aggregating their predictions when they return to create an overall forecast for each instance in the original data. Bagging is also known as bootstrap aggregation. It's a data sampling approach that uses replacement to sample data. Bootstrap aggregation is a machine learning ensemble meta-algorithm for lowering the variance of a bagged estimate, hence improving its bias and stability. Bagging classifiers combine the predictions of various estimators, reducing variance. In this study, we used five machine learning models, resulting in a total of 25 poor learners. Finally, the Bagging classifier is used, and the ensemble model's final class prediction is the class predicted by the weak learners.

2.2 Machine Learning Models

Five categories of machine learning models are used in this work, which is outlined below:

2.2.1 Logistic Regression

It's a probabilistic analytic algorithm that predicts outcomes. A more sophisticated cost function is used in Logistic Regression. The 'Sigmoid function' or 'logistic function' can be used to describe this cost function. Value of cost function should be confined

between 0 and 1, which is the rule in the logistic regression hypothesis. In classification and regression, this approach is utilized.

2.2.2 Decision Tree Classifier

A tree-structured classifier is a decision tree classifier. The internal nodes of this classifier represent attributes, branches represent decision rules, and the output is by leaf nodes. For predicting the dataset's class, the decision tree classifier starts at the root node. It compares the value of the root node with the attribute and jumps to the next node based on the result of the comparison. The technique is repeated until the tree's leaf node is reached.

2.2.3 Support Vector Classifier

In machine learning, the "Support Vector Machine" is the most commonly used Supervised Learning approach. Both classification and regression analysis can be done using this model. Optimum line or decision boundary was generated by SVM which is used to divide the n-dimensional space into classes and new data points are classified in the future. The newly generated boundary is referred to as a hyperplane. The hyperplane is created using SVM, which selects the extreme points or vectors. Support vectors are the names given to these extreme points, and the process is known as a Support Vector Machine.

2.2.4 KNN Classifier

The simplest Machine Learning algorithm is "K-Nearest Neighbour". The value of "k" has a significant impact on the correctness of the algorithm's output. KNN calculates the "Euclidean", "Manhattan", or "Minkowski" distance between feature points to compare unclassified and classified data. It is also known as a lazy learner.

2.2.5 Naïve Bayes Classifier

This machine learning model is based on the Bayes theorem and assumes predictor independence. According to the "Naïve Bayes" model, existing feature presence in a class is considered to be independent of the presence of any other feature. To develop models with analytical skills, the Naive Bayes model is used. It offers novel approaches to analyzing and comprehending datasets. When data is high, qualities are unrelated to one another, and a more efficient output is expected, Nave Bayes is chosen compared to other methods.

In the proposed method, five weak learners such as "Logistic Regression Model", "Decision Tree", "Support Vector Machine", "K-Nearest Neighbor", and "Naive Bayes" are used. We used five machine learning models in this investigation, resulting in a total of 25 weak learners. Finally, the Bagging classifier is used, and the final class prediction of the ensemble model is the class predicted by the weak learners. The accuracy of each model is determined and compared to the "Hybrid Ensemble Model", which is a composite of all five models.

3 Results and Discussion

The accuracy of the machine learning models such as "Logistic Regression Model", "Decision Tree", "Support Vector Machine", "K-Nearest Neighbor", "Naive Bayes", and "Hybrid Ensemble model" is shown in Figure 6.

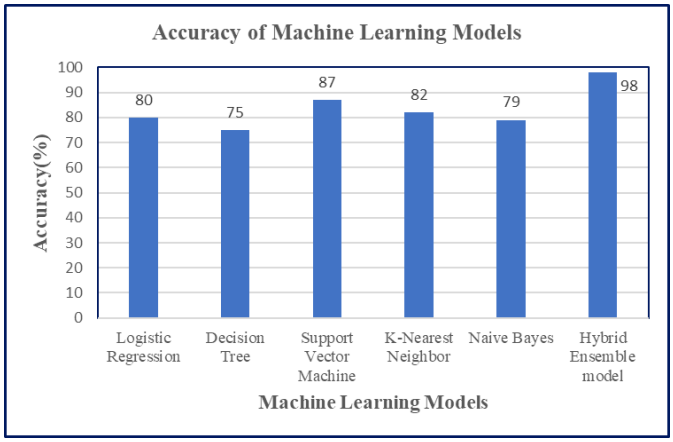

Fig 6. Accuracy of Machine Learning Models

Figure 6 demonstrates that Logistic Regression has an accuracy of 80%, Decision Tree has an accuracy of 75%, SVM has an accuracy of 87%, KNN has an accuracy of 82%, Nave Bayes has an accuracy of 79 % and the proposed Hybrid Ensemble model has an accuracy of 98%. With 98 % accuracy, the "Hybrid Ensemble model" surpassed all of the individual models, allowing the physician to effectively identify heart disease.

The accuracy comparison of the heart dataset for predicting heart disease by various authors and the proposed "Hybrid Ensemble model" is shown in Table 2.

Table 2. Accuracy comparison on the heart dataset by various authors with the proposed model

Author	Techniques	Accuracy (%)
Chu-Hsing Lin et al.[1]	Convolutional Neural Networks (CNN)	93
Montu Saw et al.[4]	Logistic Regression	80
Noor Basha et al.[5]	KNN	85
Halima El Hamdaoui et al.[8]	Naïve Bayes	84.28
Rohit Bharti et al.[10]	Deep Learning	94.28
Muhammad Zeeshan Younas[12]	Logistic Regression	86.89
Abdulwahab Ali Almazroi et al.[13]	Decision Tree	82
D. Deepika et al.[14]	MLP-EBMDA	94.28
Abdulaziz Albahr et al.[15]	RSD-ANN	96.3
Proposed	**"Hybrid Ensemble Model"**	**98**

Table 2 shows that the proposed "Hybrid Ensemble Model" has the highest accuracyof 98 percent compared with the previous works done by various authors.

In[4], the sigmoid function was used in a Logistic prediction model to predict heart disease with an accuracy of 80%.

Noor Basha et al.[5] analysis was done on KNN, NB, SVM, DT, and RF; and results were found that KNN achieved the highest accuracy of 85 percent.

Halima El Hamdaoui et al.[8] used split and cross-validation approaches in "NB", "KNN", "SVM", "RF", and "DT" methodologies to test the accuracy of heart disease prediction. NB achieved the highest accuracy of 84.28 percent using the split data technique.

In[10], researchers used three different techniques for RF, LR, KNN, SVM, DT, and XGBoost; among machine learning models, KNN had the highest accuracy of 84.86 percent, while deep learning algorithm had the highest accuracy of 94.28 percent.

Muhammad Zeeshan[12] used machine learning and data mining approaches to discover that LR outperforms other models in predicting cardiac disease, with an accuracy of 86.89 percent.

Abdulwahab's[13] study states that ANN done with the least performance and Decision tree is better than the LR, SVM, and ANN. The decision tree with a maximum depth of 4 reached the highest accuracy of 80%.

Deepika et al.[14] proposed a novel hybrid approach MLP-EBMDA and achieved the highest accuracy of 94.28% in predicting heart disease.

Abdulaziz Albahr et al.[15] provided a novel computational strategy based on a new regularizer, and testing showed that the RSD-ANN technique obtained an average validation accuracy of 96.3 percent utilizing holdout or tenfold cross-validation methods.

As mentioned in Table 2, our proposed "Hybrid Ensemble Model" has the highest accuracy of 98 percent compared with the previous works done in[1,4,5,8,10,13–15]

4 Conclusion and future scope

We proposed a "Hybrid Ensemble Model" in this study, in which we compared the accuracy of weak learners such as "Logistic Regression," "Decision Tree," "Support Vector Machine", "K-Nearest Neighbor", and "Naive Bayes" to the proposed "Hybrid Ensemble Model," which yielded encouraging results. Many researchers have previously suggested in various studies to apply the machine learning techniques and achieved higher results. Our proposed model, on the other hand, predicted the best with 98% accuracy, and thus this study might be useful to doctors and patients in predicting heart disease in advance.

As the scope of future work, this research can be extended to larger datasets, comparing the proposed technique with deep learning models. Various alternative optimization approaches, as well as different methods of data normalization, can be applied,

and the results may be compared to improve accuracy. Incorporating the proposed model with user-friendly mobile or web-based application can be developed for the easier usage of doctors and patients in the real world.

References

1) Lin CH, Yang PK, Lin YC, Fu PK. On Machine Learning Models for Heart Disease Diagnosis. *Second IEEE Eurasia Conference on Biomedical Engineering, Healthcare and Sustainability*. 2020. doi:10.1109/ECBIOS50299.2020.9203614.
2) Zeinulla E, Bekbayeva K, Yazici A. Effective diagnosis of heart disease imposed by incomplete data based on fuzzy random forest. *2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. 2020;p. 2020–2020. doi:10.1109/FUZZ48607.2020.9177531.
3) Kasabe R, Narang G. Heart Disease Prediction using Machine Learning. *International Journal of Engineering Research & Technology (IJERT)*. 2020. Available from: http://dx.doi.org/10.17577/IJERTV9IS080128.
4) Saw M, Saxena T, Kaithwas S, Yadav R, Lal N. Estimation of Prediction for Getting Heart Disease Using Logistic Regression Model of Machine Learning. *International Conference on Computer Communication and Informatics (ICCCI)*. 2020. doi:10.1109/ICCCI48352.2020.9104210.
5) Basha N, Kumar PSA, Krishna CG, Venkatesh P. Early Detection of Heart Syndrome Using Machine Learning Technique. *4th International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques (ICEECCOT)*. 2019. doi:10.1109/ICEECCOT46775.2019.9114651.
6) Katarya R, Srinivas P. Predicting Heart Disease at Early Stages using Machine Learning: A Survey. *International Conference on Electronics and Sustainable Communication Systems (ICESC)*. 2020. doi:10.1109/ICESC48915.2020.9155586.
7) Miao L, Guo X, Abbas HT, Qaraqe KA, Abbasi QH. Using Machine Learning to Predict the Future Development of Disease. *2020 International Conference on UK-China Emerging Technologies (UCET)*. 2020. doi:10.1109/UCET51115.2020.9205373.
8) Halima EL, Hamdaoui, Saïd B, Houda NE, Mustapha C, Maaroufi A. A Clinical support System for Prediction of Heart Disease using Machine Learning Techniques. *5th International Conference on Advanced Technologies for Signal and Image Processing*. 2020. doi:10.1109/ATSIP49331.2020.9231760.
9) Ananey-Obiri D, Sarku E. Predicting the Presence of Heart Diseases using Comparative Data Mining and Machine Learning Algorithms. *International Journal of Computer Applications*. 2020;176(11):17–21. Available from: https://dx.doi.org/10.5120/ijca2020920034.
10) Bharti R, Khamparia A, Shabaz M, Dhiman G, Pande S, Singh P. Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning. *Computational Intelligence and Neuroscience*. 2021;2021:1–11. Available from: https://dx.doi.org/10.1155/2021/8387680.
11) Srivenkatesh M. Prediction of Cardiovascular Disease using Machine Learning Algorithms. *International Journal of Engineering and Advanced Technology*. 2020;9(3):2404–2414. doi:10.35940/ijeat.B3986.029320.
12) Zeeshan YM. Effective Heart Disease Prediction using Machine Learning and DataMining Techniques. *International Research Journal Of Engineering And Technology (IRJET)*. 2021;8:3539–3546. Available from: ww.irjet.net.
13) Ali A, Almazroi. Survival prediction among heart patients using machine learning techniques. *Mathematical Biosciences and Engineering*. 2022;19(1):134–145. doi:10.3934/mbe.2022007.
14) Deepika D, Balaji N. Effective heart disease prediction using novel MLP-EBMDA approach. *Biomedical Signal Processing and Control*. 2022;72:103318–103318. Available from: https://dx.doi.org/10.1016/j.bspc.2021.103318.
15) Albahr A, Albahar M, Thanoon M, Binsawad M. Computational Learning Model for Prediction of Heart Disease Using Machine Learning Based on a New Regularizer. *Computational Intelligence and Neuroscience*. 2021;2021:1–10. Available from: https://dx.doi.org/10.1155/2021/8628335.