Journal of Cardiovascular Disease Research

ISSN: 0975-3583, 0976-2833 VOL12, ISSUE01, 2021

Random Forest Classifier based on Heart Disease Prediction

Sharath Pokala, Bandari Nithya

Department of Electronics and Communication Engineering

Sree Dattha Group of Institutions, Hyderabad, Telangana, India.

Abstract

In the medical field, the diagnosis of heart disease is the most difficult task. The diagnosis of heart disease is difficult as a decision relied on grouping of large clinical and pathological data. Due to this complication, the interest increased in a significant amount between the researchers and clinical professionals about the efficient and accurate heart disease prediction. In case of heart disease, the correct diagnosis in early stage is important as time is the especially important factor. Heart disease is the principal source of deaths widespread, and the prediction of heart disease is significant at an untimely phase. Machine learning in recent years has been the evolving, reliable, and supporting tools in medical domain and has provided the greatest support for predicting disease with correct case of training and testing.

Keywords: Heart disease, machine learning, random forest algorithm.

1. Introduction

It is difficult to identify heart disease because of several contributory risk factors such as diabetes, high blood pressure, high cholesterol, abnormal pulse rate and many other factors. Heart disease is a collection of diseases impacting the heart and veins of human beings. Cardiac disease symptoms vary depending on the specific type of cardiac disease. Detecting and diagnosing the cardiovascular disease is an on-going job that can be achieved with enough experience and knowledge by a qualified professional. There are many factors including age, diabetes, smoking, overweight, junk foods diet and so on. Several factors/parameters have been identified that cause heart disease or increase cardiac disease. Most hospitals have management software for monitoring their clinical and/or patient data. It is popular now and Such systems produce enormous amounts of patient information. These data are seldom used for clinical decision-making support. These data are valuable and information is kept largely unused in these data. It is an extremely difficult task to turn the accumulated clinical data into useful information that can make intelligent systems support decision-making for healthcare practitioners. This factor led to research on the processing of medical pictures Due to the lack of experts and the number of cases incorrectly diagnosed, a rapid and efficient automated detection system was required. The main purpose is to classify the key features of the medical data using the classifier model and use the models for the early prediction of cardiac disease.

According to WHO, Heart Diseases are a leading cause of death worldwide. It is quite difficult to identify the cardiovascular disease (CVD) because of some contributory factors which contribute to CVD like high blood pressure, cholesterol level, diabetics, abnormal pulse rate, and many other factors. Sometimes CVD symptoms may vary for different genders. For example, a male patient is more likely to have chest pain while a female patient has some other symptoms with chest pain like chest discomfort: such as nausea, extreme fatigue, and shortness of breath. Researchers have been exploring a wide range of techniques to predict heart diseases but the disease prediction at an early stage is not very efficient due to many factors, including but not limited to complexity, execution time, and accuracy of the approach. As such, proper treatment and diagnosis can save many lives.

One American dies every 36 seconds due to CVD. More than.665 million people die due to heart disease which 1 in every 4 deaths. Cardiovascular disease costs a lot to the US healthcare system. In

ISSN: 0975-3583, 0976-2833 VOL12, ISSUE01, 2021

the years 2014 and 2015, it cost about $219 billion per year in terms of healthcare services, medicine, and lost productivity due to death. Early diagnosis can also help to prevent heart failure which can lead to the death of a person. Angiography is considered as the most precise and accurate method for the prediction of cardiac artery disease (CAD), but it is very costly which makes it less accessible to low-income families. Rani et. al [1] proposed the authors have a hybrid decision support system that can assist in the early detection of heart disease based on the clinical parameters of the patient. Authors have used multivariate imputation by chained equations algorithm to handle the missing values. A hybridized feature selection algorithm combining the Genetic Algorithm (GA) and recursive feature elimination has been used for the selection of suitable features from the available dataset. Further for pre-processing of data, SMOTE (Synthetic Minority Oversampling Technique) and standard scalar methods have been used. In the last step of the development of the proposed hybrid system, authors have used support vector machine, naive bayes, logistic regression, random forest, and Adaboost classifiers. It has been found that the system has given the most accurate results with random forest classifier. The proposed hybrid system was tested in the simulation environment developed using Python. It was tested on the Cleveland heart disease dataset available at UCI (University of California, Irvine) machine learning repository. It has achieved an accuracy of 86.6%, which is superior to some of the existing heart disease prediction systems found in the literature.

2. Literature survey

Shah et. al [4] presents various attributes related to heart disease, and the model on basis of supervised learning algorithms as Naïve Bayes, decision tree, K-nearest neighbor, and random forest algorithm. It uses the existing dataset from the Cleveland database of UCI repository of heart disease patients. The dataset comprises 303 instances and 76 attributes. Of these 76 attributes, only 14 attributes are considered for testing, important to substantiate the performance of different algorithms. This research paper aims to envision the probability of developing heart disease in the patients. The results portray that the highest accuracy score is achieved with K-nearest neighbor.

Guoet. Al [5] proposed Recursion enhanced random forest with an improved linear model (RFRF-ILM) to detect heart disease. This paper aims to find the key features of the prediction of cardiovascular diseases through the use of machine learning techniques. The prediction model is adding various combinations of features and various established methods of classification. it produces a better level of performance with precision through the heart disease prediction model.

Hager Ahmed et. al [6] presented a system for real-time heart disease prediction that was developed based on Apache Spark and Apache Kafka. Our real-time system consists of three components, namely Offline Model Building, Stream Processing Pipeline, and Online Prediction. Offline Model Building is a machine learning model that can achieve high accuracy. In this component, they analyze the features in the dataset and select the optimal set of features based on two feature selection algorithms.

Katarya et. al [7] discussed the heart disease and its risk factors and explained machine learning techniques. Using that machine learning techniques, they have predicted heart disease and provided a comparative analysis of the algorithms for machine learning used for the experiment of the prediction

Kannan et. al [8] examined and compared the accuracy of four different machine learning algorithms with receiver operating characteristic (ROC) curve for predicting and diagnosing heart disease by the 14 attributes from UCI Cardiac Datasets.

ISSN: 0975-3583, 0976-2833 VOL12, ISSUE01, 2021

Mamun Ali et. al [9] found that using a heart disease dataset collected from Kaggle three-classification based on k-nearest neighbor (KNN), decision tree (DT) and random forests (RF) algorithms the RF method achieved 100% accuracy along with 100% sensitivity and specificity. Thus, they found that a relatively simple supervised machine learning algorithm can be used to make heart disease predictions with very high accuracy and excellent potential utility.

3. Proposed system

Dataset description

14-Attributes

303-Rows

15-Columns

Columns description

age: age in years

sex: (1 = male; 0 = female)

cp: chest pain type

trestbps: resting blood pressure (in mm Hg on admission to the hospital)

chol: serum cholestoral in mg/dl

fbs: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)

restecg: resting electrocardiographic results

thalach: maximum heart rate achieved

exang: exercise induced angina (1 = yes; 0 = no)

oldpeak: ST depression induced by exercise relative to rest

slope: the slope of the peak exercise ST segment

ca: number of major vessels (0-3) colored by flourosopy

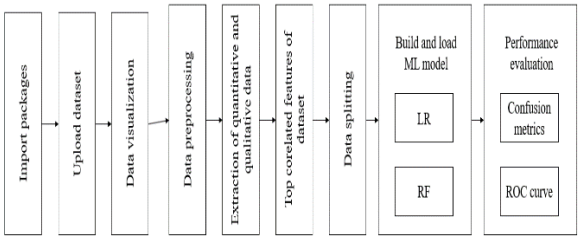
thal: 3 = normal; 6 = fixed defect; 7 = reversable defect

target: refers to the presence of heart disease in the patient (1=yes, 0=no)

0, 63, 1, 3, 145, 233, 1, 0, 150, 0, 2.3, 0, 0, 1, 1

1, 37, 1, 2, 130, 250, 0, 1, 187, 0, 3.5, 0, 0, 2, 1

2, 41, 0, 1, 130, 204, 0, 0, 172, 0, 1.4, 2, 0, 2, 1



ISSN: 0975-3583, 0976-2833 VOL12, ISSUE01, 2021

Fig. 1: Block diagram of proposed system.

Data Preprocessing in Machine learning

Data pre-processing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model.

When creating a machine learning project, it is not always a case that we come across the clean and formatted data. And while doing any operation with data, it is mandatory to clean it and put in a formatted way. So, for this, we use data pre-processing task.

Why do we need Data Pre-processing?

A real-world data generally contains noises, missing values, and maybe in an unusable format which cannot be directly used for machine learning models. Data pre-processing is required tasks for cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine learning model.

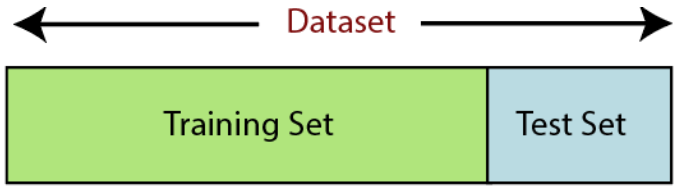
- Getting the dataset
- Importing libraries
- Importing datasets
- Finding Missing Data
- Encoding Categorical Data
- Splitting dataset into training and test set
- Feature scaling

Splitting the Dataset into the Training set and Test set

In machine learning data pre-processing, we divide our dataset into a training set and test set. This is one of the crucial steps of data pre-processing as by doing this, we can enhance the performance of our machine learning model.

Suppose if we have given training to our machine learning model by a dataset and we test it by a completely different dataset. Then, it will create difficulties for our model to understand the correlations between the models.

If we train our model very well and its training accuracy is also very high, but we provide a new dataset to it, then it will decrease the performance. So we always try to make a machine learning model which performs well with the training set and also with the test dataset. Here, we can define these datasets as:



Training Set: A subset of dataset to train the machine learning model, and we already know the output.

Test set: A subset of dataset to test the machine learning model, and by using the test set, model predicts the output.

ISSN: 0975-3583, 0976-2833 VOL12, ISSUE01, 2021

Random Forest Algorithm

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

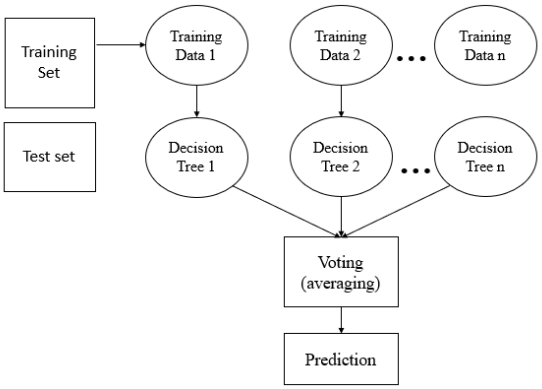


Fig. 2: Random Forest algorithm.

Random Forest algorithm

Step 1: In Random Forest n number of random records are taken from the data set having k number of records.

Step 2: Individual decision trees are constructed for each sample.

Step 3: Each decision tree will generate an output.

Step 4: Final output is considered based on Majority Voting or Averaging for Classification and regression respectively.

Important Features of Random Forest

- **Diversity**- Not all attributes/variables/features are considered while making an individual tree, each tree is different.
- **Immune to the curse of dimensionality**- Since each tree does not consider all the features, the feature space is reduced.
- **Parallelization**-Each tree is created independently out of different data and attributes. This means that we can make full use of the CPU to build random forests.
- **Train-Test split**- In a random forest we don't have to segregate the data for train and test as there will always be 30% of the data which is not seen by the decision tree.
- **Stability**- Stability arises because the result is based on majority voting/ averaging.

Assumptions for Random Forest

ISSN: 0975-3583, 0976-2833 VOL12, ISSUE01, 2021

Since the random forest combines multiple trees to predict the class of the dataset, it is possible that some decision trees may predict the correct output, while others may not. But together, all the trees predict the correct output. Therefore, below are two assumptions for a better Random Forest classifier:

- There should be some actual values in the feature variable of the dataset so that the classifier can predict accurate results rather than a guessed result.
- The predictions from each tree must have very low correlations.

Below are some points that explain why we should use the Random Forest algorithm

- It takes less training time as compared to other algorithms.
- It predicts output with high accuracy, even for the large dataset it runs efficiently.
- It can also maintain accuracy when a large proportion of data is missing.

Types of Ensembles

Before understanding the working of the random forest, we must look into the ensemble technique. Ensemble simply means combining multiple models. Thus, a collection of models is used to make predictions rather than an individual model. Ensemble uses two types of methods:

Bagging– It creates a different training subset from sample training data with replacement & the final output is based on majority voting. For example, Random Forest. Bagging, also known as Bootstrap Aggregation is the ensemble technique used by random forest. Bagging chooses a random sample from the data set. Hence each model is generated from the samples (Bootstrap Samples) provided by the Original Data with replacement known as row sampling. This step of row sampling with replacement is called bootstrap. Now each model is trained independently which generates results. The final output is based on majority voting after combining the results of all models. This step which involves combining all the results and generating output based on majority voting is known as aggregation.

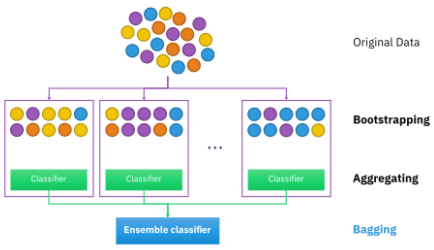


Fig. 3: RF Classifier analysis.

Boosting– It combines weak learners into strong learners by creating sequential models such that the final model has the highest accuracy. For example, ADA BOOST, XG BOOST.

ISSN: 0975-3583, 0976-2833 VOL12, ISSUE01, 2021

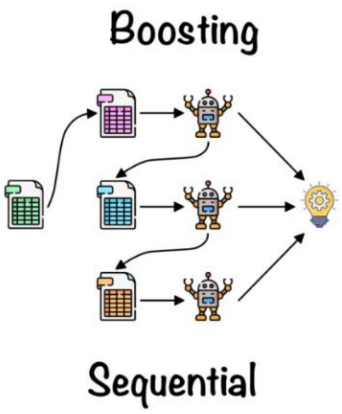


Fig. 4: Boosting RF Classifier.

Advantages of proposed system

- It can be used in classification and regression problems.
- It solves the problem of overfitting as output is based on majority voting or averaging.
- It performs well even if the data contains null/missing values.
- Each decision tree created is independent of the other thus it shows the property of parallelization.
- It is highly stable as the average answers given by a large number of trees are taken.
- It maintains diversity as all the attributes are not considered while making each decision tree though it is not true in all cases.
- It is immune to the curse of dimensionality. Since each tree does not consider all the attributes, feature space is reduced.

4. Results

Sample Dataset

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	
count	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.0
mean	54.366337	0.683168	0.966997	131.623762	246.264026	0.148515	0.528053	149.646865	0.326733	1.039604	1.399340	0.729373	2.3
std	9.082101	0.466011	1.032052	17.538143	51.830751	0.356198	0.525860	22.905161	0.469794	1.161075	0.616226	1.022606	0.6
min	29.000000	0.000000	0.000000	94.000000	126.000000	0.000000	0.000000	71.000000	0.000000	0.000000	0.000000	0.000000	0.0
25%	47.500000	0.000000	0.000000	120.000000	211.000000	0.000000	0.000000	133.500000	0.000000	0.000000	1.000000	0.000000	2.0
50%	55.000000	1.000000	1.000000	130.000000	240.000000	0.000000	1.000000	153.000000	0.000000	0.800000	1.000000	0.000000	2.0
75%	61.000000	1.000000	2.000000	140.000000	274.500000	0.000000	1.000000	166.000000	1.000000	1.600000	2.000000	1.000000	3.0
max	77.000000	1.000000	3.000000	200.000000	564.000000	1.000000	2.000000	202.000000	1.000000	6.200000	2.000000	4.000000	3.0

The scale of each feature column is different and quite varied as well. While the maximum for age reaches 77, the maximum of chol (serum cholestoral) is 564

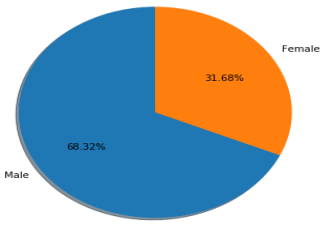
Column renaming

```
Index(['age', 'sex', 'chest_pain_type', 'resting_blood_pressure',
       'cholesterol', 'fasting_blood_sugar', 'rest_ecg',
       'max_heart_rate_achieved', 'exercise_induced_angina', 'st_depression',
       'st_slope', 'num_major_vessels', 'thalassemia', 'target'],
      dtype='object')
```

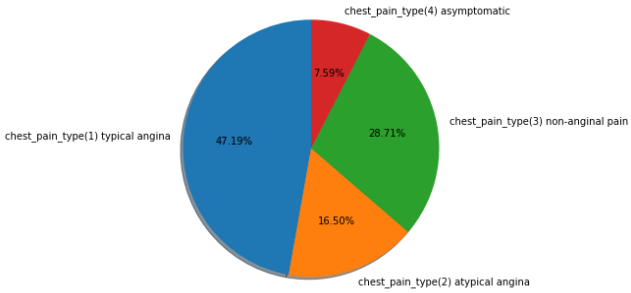
Count of each target class

ISSN: 0975-3583, 0976-2833 VOL12, ISSUE01, 2021

Male vs Female data

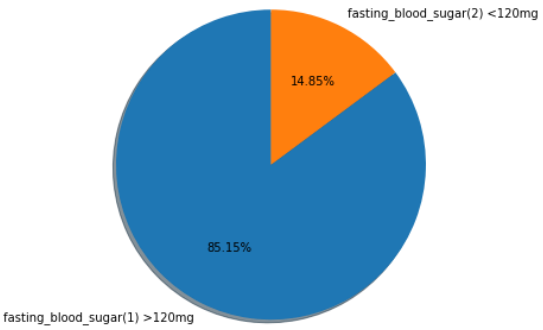


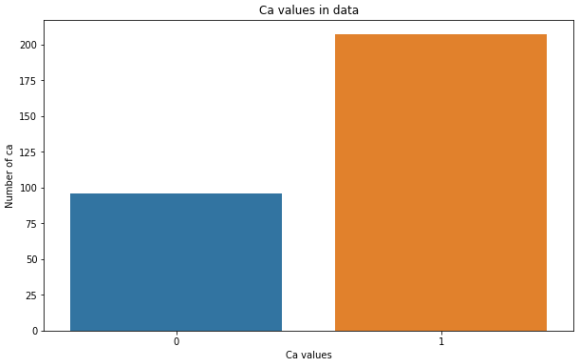
Chest pain type



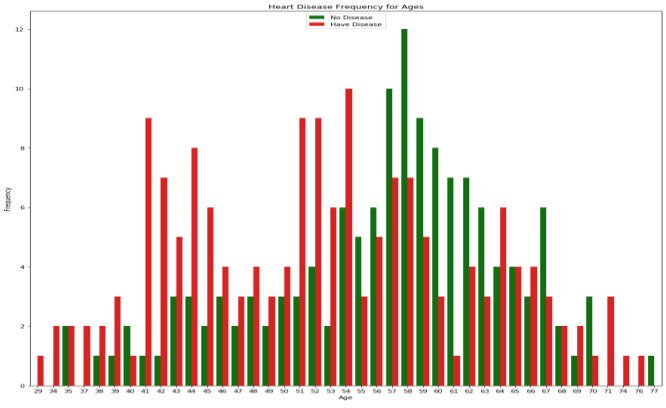
Fasting Blood Sugar

ISSN: 0975-3583, 0976-2833 VOL12, ISSUE01, 2021

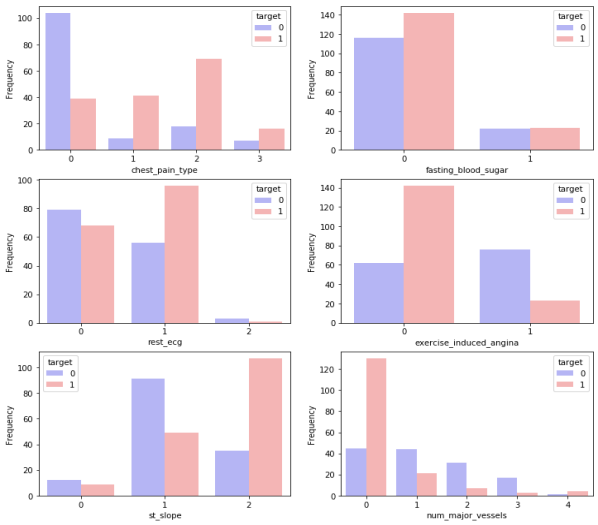




Heart disease frequency by age

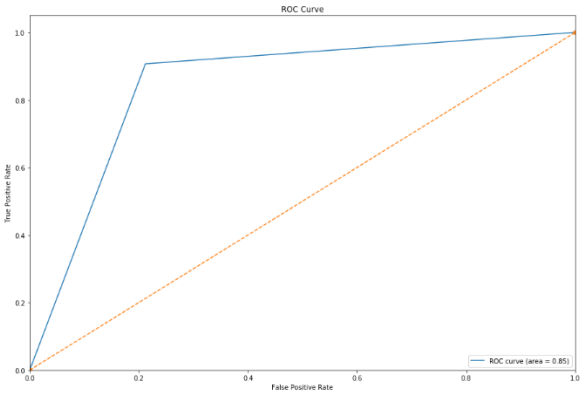


ISSN: 0975-3583, 0976-2833 VOL12, ISSUE01, 2021



Classification report of Random Forest

	precision	recall	f1-score	support
0	0.87	0.79	0.83	33
1	0.85	0.91	0.88	43
accuracy			0.86	76
macro avg	0.86	0.85	0.85	76
weighted avg	0.86	0.86	0.85	76



Prediction on test data

	patient_id	Heart_Disease
0	0	N
1	1	N
2	2	N
3	3	N
4	4	N
5	5	N
6	6	N
7	7	Y
8	8	Y
9	9	N
10	10	N
11	11	Y
12	12	N
13	13	N
14	14	N
15	15	N

ISSN: 0975-3583, 0976-2833 VOL12, ISSUE01, 2021

5. Conclusion

Identifying the processing of raw healthcare data of heart information will help in the long-term saving of human lives and early detection of abnormalities in heart conditions. Machine learning techniques were used in this work to process raw data and provide a new and novel discernment towards heart disease. Heart disease prediction is challenging and very important in the medical field. However, the mortality rate can be drastically controlled if the disease is detected at the early stages and preventative measures are adopted as soon as possible. Further extension of this study is highly desirable to direct the investigations to real-world datasets instead of just theoretical approaches and simulations. The proposed hybrid HRFLM approach is used combining the characteristics of Random Forest (RF).

5.1 Future scope

The future course of this work can be performed with diverse mixtures of machine learning techniques to better prediction techniques. Furthermore, new feature-selection methods can be developed to get a broader perception of the significant features to increase the performance of heart disease prediction.

References

[1] Rani, P., Kumar, R., Ahmed, N.M.O.S. et al. A decision support system for heart disease prediction based upon machine learning. J Reliable Intell Environ 7, 263–275 (2021). https://doi.org/10.1007/s40860-021-00133-6

[2] M. Kavitha, G. Gnaneswar, R. Dinesh, Y. R. Sai and R. S. Suraj, "Heart Disease Prediction using Hybrid machine Learning Model," 2021 6th International Conference on Inventive Computation Technologies (ICICT), 2021, pp. 1329-1333, doi: 10.1109/ICICT50816.2021.9358597.

[3] 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.

[4] Shah, D., Patel, S. & Bharti, S.K. Heart Disease Prediction using Machine Learning Techniques. SN COMPUT. SCI. 1, 345 (2020). https://doi.org/10.1007/s42979-020-00365-y

[5] C. Guo, J. Zhang, Y. Liu, Y. Xie, Z. Han and J. Yu, "Recursion Enhanced Random Forest with an Improved Linear Model (RERF-ILM) for Heart Disease Detection on the Internet of Medical Things Platform," in IEEE Access, vol. 8, pp. 59247-59256, 2020, doi: 10.1109/ACCESS.2020.2981159.

[6] Hager Ahmed, Eman M.G. Younis, Abdeltawab Hendawi, Abdelmgeid A. Ali, Heart disease identification from patients' social posts, machine learning solution on Spark, Future Generation Computer Systems, Volume 111, 2020, Pages 714-722, ISSN 0167-739X, https://doi.org/10.1016/j.future.2019.09.056.

[7] Katarya, R., Meena, S.K. Machine Learning Techniques for Heart Disease Prediction: A Comparative Study and Analysis. Health Technol. 11, 87–97 (2021). https://doi.org/10.1007/s12553-020-00505-7

[8] Kannan, R., Vasanthi, V. (2019). Machine Learning Algorithms with ROC Curve for Predicting and Diagnosing the heart disease. In: Soft Computing and Medical Bioinformatics.

ISSN: 0975-3583, 0976-2833 VOL12, ISSUE01, 2021

SpringerBriefs in Applied Sciences and Technology (). Springer, Singapore. https://doi.org/10.1007/978-981-13-0059-2_8

[9] Md Mamun Ali, Bikash Kumar Paul, Kawsar Ahmed, Francis M. Bui, Julian M.W. Quinn, Mohammad Ali Moni, Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison, Computers in Biology and Medicine, Volume 136, 2021, 104672, ISSN 0010-4825, https://doi.org/10.1016/j.compbiomed.2021.104672.