IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

Impact Factor: 7.301

Volume 3, Issue 16, May 2023

Cardiovascular Disease Prediction using Deep Learning and Feature Selection

Shrivatsa S. Desai[1], Kunal S. Gajmal[2], Suraj S. Bhosale[3], Aniket B. Manjare[4]

Department of Computer Engineering
RMD Sinhgad School of Engineering, Warje, Pune, India
Savitribai Phule Pune University, Pune, India

Abstract*:* *Due to a variety of alterations in human lifestyles, cardiovascular disease is one of the primary causes of death worldwide. If diagnosed early enough, heart disease can be minimised in about 90% of cases, giving doctors valuable insight about how to diagnose and treat patients. One of the best methods for making predictions is the use of machine learning. Studies on applying ML systems to forecast heart disease only look at the broad picture. Predicting the disease and its root cause is one of the toughest problems we face today. With the use of deep learning algorithms, we have developed an innovative approach in this study to recognise big datasets, improving the precision of cardiovascular disease prediction.In our model, feature selection and artificial neural networks have been used to predict cardiovascular diseases. Feature selection and ANN are two methods based on machine learning (ML) that can be used to select the most pertinent features from a dataset and give helpful prediction results. The accuracy of the two models, which are applied to analyse two distinct datasets, is 83% and 97.42%, respectively.*

Keywords: Artificial Neural Networks(ANN), Feature Selection, Cardiovascular Disease, Prediction

I. INTRODUCTION

Cardiovascular disease is a term that refers to numerous kinds of diseases related to the heart. Heart disease is another term for cardiovascular disease (CVD). Cardiovascular disease has been the most serious cause of death globally over the past ten years. For the patient to be given the most effective treatment, the diagnostic system has to deliver accurate results. The root cause of the sickness was discovered using a variety of previous methods. There are two separate categories for elements that could be changed and elements that could not be changed. Age, smoking, inheritance, sex, high blood pressure, eating poorly, drinking alcohol, and not exercising enough are all contributors to the risk of heart disease, although the results were not as precise.

Machine learning proves to be effective in making choices and predictions from the vast amounts of data generated by the healthcare sector. This research aims to predict potential cardiovascular disease by analyzing the data of patients to categorize whether they have heart disease or not using a machine-learning algorithm. Machine learning methods can be extremely beneficial in such circumstances. There is a common set of basic risk factors that determine whether or not someone will ultimately be at risk for heart disease, regardless of the fact that heart disease can manifest itself in a number of ways.It is now important to develop a new system that can predict cardiovascular issues in an easy and less expensive manner due to the rising number of heart diseases.

The model of methods like artificial neural networks and feature selection makes up the proposed system. The ANN will assist in creating a model, initialising the weights, and extracting the crucial features from the trained dataset. Additionally, two separate datasets have been used to test the proposed method.

The main objectives of this system are:

- The aim of the model is to predict if the patient will be diagnosed with cardiovascular disease or not based on the binary outcome. So if the result is 1, then the patient will be diagnosed with cardiovascular disease, and if it is 0, then the patient will not be diagnosed with cardiovascular disease.
- The prediction should be done with the minimum number of attributes possible.

DOI: 10.48175/IJARSCT-10972

II. MATERIALS AND METHODS

2.1 Dataset Description

The name of first dataset is Heart Disease Dataset taken from Kaggle platform. The dataset has 303 instance and 14 attribute.

Table1: Description of attributes of dataset.

Sr.No	Feature Name	Description of Features	Values
1	AGE	Age of the patient in years	-
2	SEX	Gender of patient	1=male 0=female
3	CP	Type of chest pain	0=Atypical angina, 1=typical angina , 2=asymptotic, 3=non angina pain
4	TRESTBPS	Resting Blood pressure	94-200
5	CHOL	Serum cholesterol level	126-564
6	FBS	Fasting blood sugar 1>=120,0<=120	0=false 1=true
7	RESTECG	Resting electrocardiographic results	0=normal 1=ST-T wave abnormalities 2= left ventricular hypertrophy
8	THALACH	Maximum heart rate Achieved.	71-202
9	EXANG	Exercise Induced Angina	0=no 1=yes
10	OLD PEAK	ST depression induced by exercise related to rest	0.0-6.2
11	SLOPE	Slope of the peak exercise ST segment	0= un sloping 1=flat 2=down sloping
12	CA	Count of major vessels coloured By Fluoroscopy	0-3
13	THAL	Thallium Scan	3=normal 6=fixed 7=reversible effect
14	Target	Class Attribute	0=no 1=yes

The dataset contains 8 categorical attributes and 6 numeric attributes. Table 1 contains the complete information about the dataset.

The male patient has gender value of one and female patient has gender value of zero. Male patients are at the high risk of heart disease than that of female patients.

The many types of chest pain include asymptotic, non-angina discomfort, typical angina, and typical atypical angina. The chest pain known as angina is brought on by a lack of rich oxygen blood that the heart receives. Stress on the mind or emotions might lead to atypical angina. Asymptotic is not a sign of cardiovascular disease.

TRESTBPS indicates the resting blood pressure value of an individual the unit of TRESTBPS is mmHg. The TRESTBPS attribute in the dataset ranges from 94-200.

Serum Cholesterol is the total level of cholesterol accumulated which is ranging from 126-564.

FBS indicates the fasting blood sugar value of an individual. If the FBS is less than 120mg/dl then the value is 1. If the FBS is more than 120mg/dl then the value assigned to the attribute is 0.

RESTECG displays the resting electrocardiographic result the value is assigned to 0 if the RESTECG is normal, the value is assigned to 1 if the RESTECG have ST-T wave abnormality. The value is assigned to 2 if RESTECG have left ventricular hypertrophy.

THALACH represents Maximum Heart Rate Achieved by an individual. Increase in heart beat rate by 10% increase the cardiac death by at least 20%.

DOI: 10.48175/IJARSCT-10972

EXANG is Exercise Induced Angina. EXANG is recorded as 0 if there is no pain and recorded as 1if there is pain. Angina is usually felt in the centre of the chest it may even spread to both shoulders.

OLDPEAK is the length of the ST-segment depression is crucial to taken into account because a positive ECG stress test is produced by the recovery following the peak stress.

SLOPE represents the slope of the ST segment.

THAL represents duration in exercise test in minutes which displays the thalassemia.

CA is the ECG stress test which is considered as abnormal when there is down sloping ST segment depression> 1mm at 60-80ms.

Target is the class attribute it is recorded as zero for non-diseased person and one for the person suffering with heart disease.

The name of second dataset is Cardiovascular Disease Dataset taken from Kaggle platform. The dataset has 70,000 instance and 12 attribute.

Table 2: Description of attributes of dataset.

Sr.No	Feature Name	Description of Features	Type
1	age	Age of the patient in years	int(days)
2	height	Height of patient	int(cm)
3	weight	Weight of patient	int(cm)
4	gender	Gender of an individual.	Categorical
5	ap_hi (Systolic blood pressure)	Systolic blood pressure of an individual.	int
6	ap_lo(Diastolic blood pressure)	Diastolic blood pressure of an individual.	int
7	cholestrol	Cholestrol level of an individual.	Categorical
8	gluc(Glucose)	Glucose level of an individual.	Categorical
9	Smoke	Smoking habit of an individual.	Categorical
10	alco(Alcohol intake)	Alcohol intake of an individual.	Categorical
11	active(Physical activity)	Physical activity of an individual	Categorical
12	cardio(Presence or absence of cardiovascular disease)	Presence or absence of cardiovascular disease	Categorical

The dataset contains 7 categorical attributes and 5 numeric attributes. Table 2 contains the complete information about the dataset.

The age, height, weight, and gender represent the characteristics of an individual.

The ap_hi and ap_lo are the Systolic and Diastolic blood pressures of an individual. The normal ranges of this blood pressures are lesser than 120 mm Hg and 80 mm Hg respectively. If the blood pressures goes higher than this will leads to heart related problems.

Serum Cholesterol is the total level of cholesterol accumulated which is ranging from 126-564.

The glucose indicates the glucose level of an individual. If the level goes up then it can be reason for health problem. The same thing goes with smoking and alcohol intake as the more consumption will lead to serious health problems. The Cardio represents the presence or absence of the cardiovascular disease.

DOI: 10.48175/IJARSCT-10972

III. PROPOSED SYSTEM

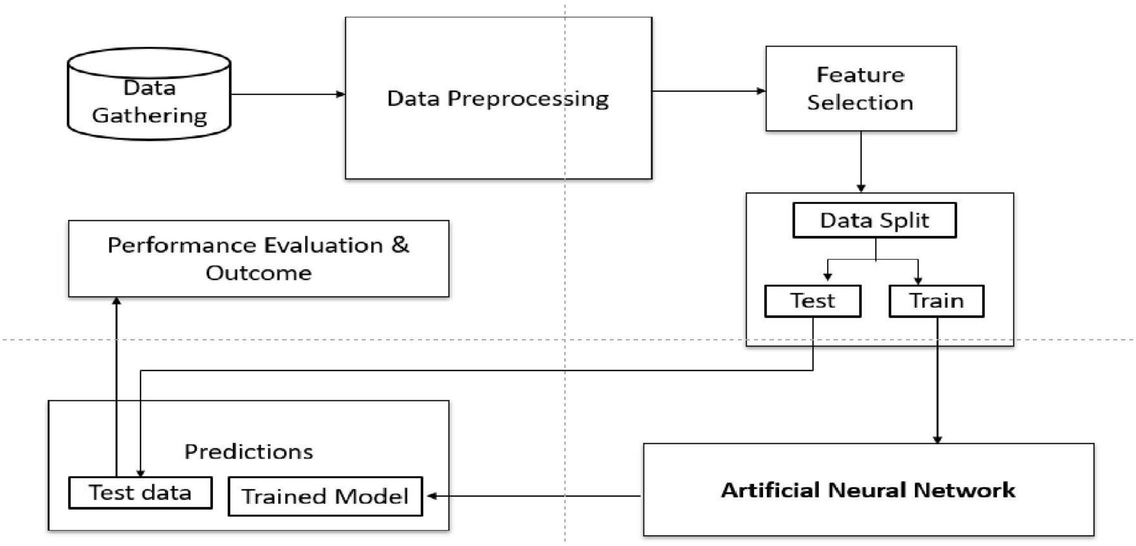


Fig1. PROPOSED SYSTEM

IV. LITERATURE SURVEY:

Sr No	Name of Journal/ Year of Published	Paper Title	Author Name	Advantages	Research Gap Identified
1.	*IEEE Access*, vol. 9, pp. 135210-135223, 2021, doi: 10.1109/ACCESS.2021.3116974.	An Efficient Prediction Method for Coronary Heart Disease Risk Based on Two Deep Neural Networks Trained on Well-Ordered Training Datasets	T. Amarbayasgalan, V. -H. Pham, N. Theera-Umpon, Y. Piao and K. H. Ryu	It works on separated, regular, and highly biassed datasets.	It is possible to improve the performance of a single predictive model trained on the whole training dataset by two different predictive models trained on the highly biased and remained common subsets
2.	*IEEE Access*, vol. 8, pp. 157643-157653, 2020, doi: 10.1109/ACCESS.2020.3015757.	Clinical Implication of Machine Learning in Predicting the Occurrence of Cardiovascular Disease Using Big Data (Nationwide Cohort Data in Korea)	G. Joo, Y. Song, H. Im and J. Park	The use of medications by the physicians provided important information on the occurrence of diseases.	To investigate a more effective ML method for the CVD risk prediction

DOI: 10.48175/IJARSCT-10972

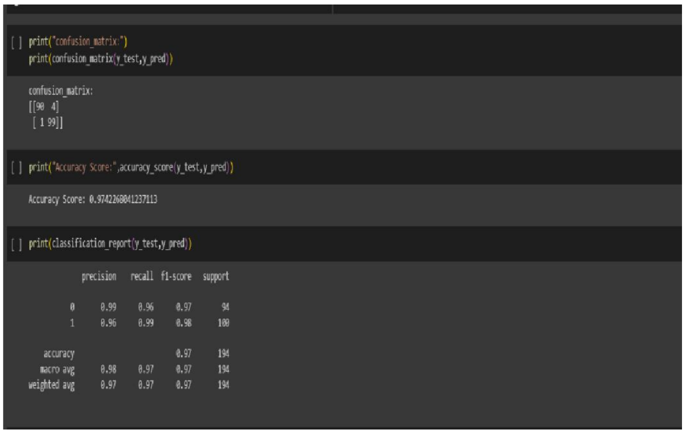
3.	*IEEE Access*, vol. 7, pp. 69559-69574, 2019, doi: 10.1109/ACCESS.2019.2912226.	Deep Ensemble Detection of Congestive Heart Failure Using Short-Term RR Intervals	L. Wang, W. Zhou, Q. Chang, J. Chen and X. Zhou	The deep ensemble models have shown improvement in accuracy.	The model has data imbalance and lack of complex techniques handling.
4.	SN Computer Science. 2. 350. 10.1007/s42979-021-00731-4.	Machine Learning Predictive Models for Coronary Artery Disease	L. J. Muhammad, Ibrahem Al-Shourbaji, Ahmed Abba Haruna,I. A. Mohammed, Abdulkadir Ahmad Muhammed Besiru Jibrin1	The model performed well on every parameter.	The decision tree generated with random forest machine learning algorithm can be converted into production rules and could be used develop expert system
5.	International Journal of Advanced Computer Science and Applications. 11. 10.14569/IJACSA.2020.0110369.	Enhanced Accuracy of Heart Disease Prediction using Machine Learning and Recurrent Neural Networks Ensemble Majority Voting Method.	Javid, Irfan & Zager, Ahmed & Ghazali, Rozaida	An increase of 2.1% in accuracy for classifiers was attained with the help of an ensemble voting-based model.	Better ensemble models of ML and DL can be made.

V. RESULT

A confusion matrix is a table that is often used to evaluate the performance of a classification model. It summarizes the predictions made by the model on a set of data and compares them with the actual labels.

A classification report is a common tool used to evaluate the performance of a classification model. It provides a detailed analysis of the model's performance by calculating various metrics such as precision, recall, F1-score, and support for each class.

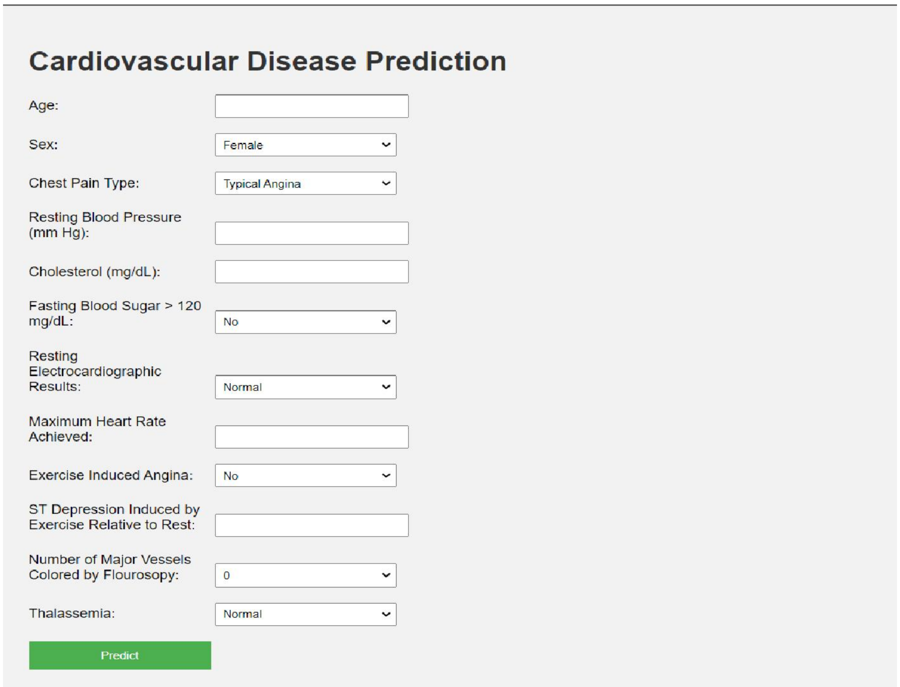
The model consists of an artificial neural network, and the feature selection shows the improved accuracy of 97.42%, in which we used the Keras tuner for the optimal hyperparameters.



```
[ ] print("confusion_matrix:")
    print(confusion_matrix(y_test,y_pred))

    confusion_matrix:
    [[90  4]
     [ 1 99]]

[ ] print("Accuracy Score:",accuracy_score(y_test,y_pred))

    Accuracy Score: 0.9742268041237113

[ ] print(classification_report(y_test,y_pred))

                  precision    recall  f1-score   support

               0       0.99      0.96      0.97        94
               1       0.96      0.99      0.98       100

        accuracy                           0.97       194
       macro avg       0.98      0.97      0.97       194
    weighted avg       0.97      0.97      0.97       194
```

Fig 1. Results of the Model

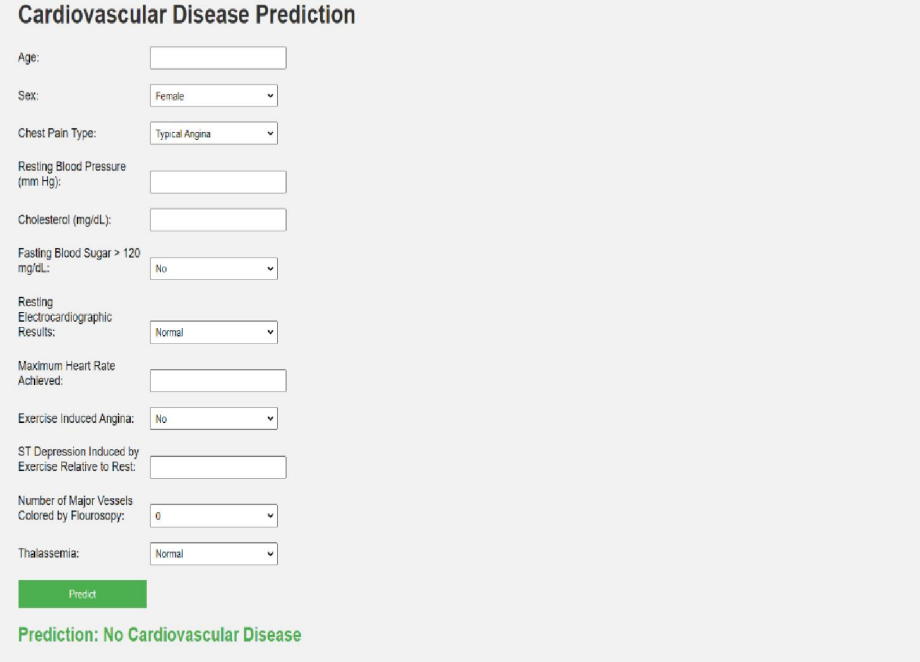
DOI: 10.48175/IJARSCT-10972



Cardiovascular Disease Prediction

Age:	
Sex:	Female
Chest Pain Type:	Typical Angina
Resting Blood Pressure (mm Hg):	
Cholesterol (mg/dL):	
Fasting Blood Sugar > 120 mg/dL:	No
Resting Electrocardiographic Results:	Normal
Maximum Heart Rate Achieved:	
Exercise Induced Angina:	No
ST Depression Induced by Exercise Relative to Rest:	
Number of Major Vessels Colored by Flourosopy:	0
Thalassemia:	Normal

Fig 2. Front End View of System



Cardiovascular Disease Prediction

Age:	
Sex:	Female
Chest Pain Type:	Typical Angina
Resting Blood Pressure (mm Hg):	
Cholesterol (mg/dL):	
Fasting Blood Sugar > 120 mg/dL:	No
Resting Electrocardiographic Results:	Normal
Maximum Heart Rate Achieved:	
Exercise Induced Angina:	No
ST Depression Induced by Exercise Relative to Rest:	
Number of Major Vessels Colored by Flourosopy:	0
Thalassemia:	Normal

Prediction: No Cardiovascular Disease

Fig 3. Predicted Front End View of System

VI. CONCLUSION

To predict cardiovascular disease and have treatment in a timely manner, early prediction is important. So, in this paper, we present a model of an artificial neural network (ANN) and feature selection to give better accuracy with the minimum number of attributes possible if there is any slight chance of having cardiovascular disease. The use of a Keras tuner provides the optimum parameters for the model. With the help of the feature selection, the model shows an improvement in accuracy. In future work, we can try to implement a deep learning model with more effective feature selection techniques, and there is also a chance of using CNN for more detailed prediction on pictorial images of patients health records.

REFERENCES

[1]Kalluri, Hemantha kumar & Tulasi Krishna, Sajja. (2020). A Deep Learning Method for Prediction of Cardiovascular Disease Using Convolutional Neural Network. Revue d intelligence artificielle. 34. 601-606. 10.18280/ria.340510.

[2]K. G. Dinesh, K. Arumugaraj, K. D. Santhosh and V. Mareeswari, "Prediction of Cardiovascular Disease Using Machine Learning Algorithms," 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT), Coimbatore, India, 2018, pp. 1-7, doi: 10.1109/ICCTCT.2018.8550857.

[3] T. Amarbayasgalan, V. -H. Pham, N. Theera-Umpon, Y. Piao and K. H. Ryu, "An Efficient Prediction Method for Coronary Heart Disease Risk Based on Two Deep Neural Networks Trained on Well-Ordered Training Datasets," in IEEE Access, vol. 9, pp. 135210-135223, 2021, doi: 10.1109/ACCESS.2021.3116974.

[4] Chaurasia, Vikas & Pal, Saurabh. (2013). Data Mining Approach to Detect Heart Diseases. International Journal of Advanced Computer Science and Information Technology (IJACSIT). 2. 56-66.

[5] N. G. B. Amma, "cardiovascular disease prediction system using genetic algorithm and neural network," 2012 International Conference on Computing, Communication and Applications, Dindigul, India, 2012, pp. 1-5, doi: 10.1109/ICCCA.2012.6179185.

[6] Dutta, Aniruddha & Batabyal, Tamal & Basu, Meheli & Acton, Scott. (2020). An Efficient Convolutional Neural Network for Coronary Heart Disease Prediction. Expert Systems with Applications. 159. 113408. 10.1016/j.eswa.2020.113408.

[7] G. Joo, Y. Song, H. Im and J. Park, "Clinical Implication of Machine Learning in Predicting the Occurrence of Cardiovascular Disease Using Big Data (Nationwide Cohort Data in Korea)," in IEEE Access, vol. 8, pp. 157643-157653, 2020, doi: 10.1109/ACCESS.2020.3015757.

[8] Rustam, Furqan & Ishaq, Abid & Munir, Kashif & Almutairi, Mubarak & Aslam, Naila & Ashraf, Imran. (2022). Incorporating CNN Features for Optimizing Performance of Ensemble Classifier for Cardiovascular Disease Prediction. Diagnostics (Basel, Switzerland). 12. 10.3390/diagnostics12061474.

[9] Patil, Prasadgouda & Pm, Mallikarjuna & P S, Ashok. (2021). Journal of Critical Reviews MACHINE LEARNING BASED ALGORITHM FOR RISK PREDICTION OF CARDIO VASCULAR DISEASE (CVD). Journal of Critical Reviews. 7. 2020. 10.31838/jcr.07.09.157.

[10]Muhammad LJ, Al-Shourbaji I, Haruna AA, Mohammed IA, Ahmad A, Jibrin MB. Machine Learning Predictive Models for Coronary Artery Disease. SN Computer Sci. 2021;2(5):350. doi: 10.1007/s42979-021-00731-4. Epub 2021 Jun 22. PMID: 34179828; PMCID: PMC8218284.

[11] Dami, Sina & Yahaghizadeh, Mahtab. (2021). Predicting cardiovascular events with deep learning approach in the context of the internet of things. Neural Computing and Applications. 33. 1-18. 10.1007/s00521-020-05542-x.

[12] A. K. Paul, P. C. Shill, M. R. I. Rabin and M. A. H. Akhand, "Genetic algorithm based fuzzy decision support system for the diagnosis of heart disease," 2016 5th International Conference on Informatics, Electronics and Vision (ICIEV), Dhaka, Bangladesh, 2016, pp. 145-150, doi: 10.1109/ICIEV.2016.7759984.

[13] Mienye, Domor & Sun, Yanxia & Wang, Zenghui. (2020). An improved ensemble learning approach for the prediction of heart disease risk. Informatics in Medicine Unlocked. 20. 100402. 10.1016/j.imu.2020.100402.

[14] Abdeldjouad, Fatma Zahra & Menaouer, Brahami & Nada, Matta. (2020). A Hybrid Approach for Heart Disease Diagnosis and Prediction Using Machine Learning Techniques. 10.1007/978-3-030-51517-1_26.

[15] Javid, Irfan & Zager, Ahmed & Ghazali, Rozaida. (2020). Enhanced Accuracy of Heart Disease Prediction using Machine Learning and Recurrent Neural Networks Ensemble Majority Voting Method. International Journal of Advanced Computer Science and Applications. 11. 10.14569/IJACSA.2020.0110369.

[16] Christalin, Beulah & Jeeva, Carolin. (2019). Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. Informatics in Medicine Unlocked. 16. 100203. 10.1016/j.imu.2019.100203.

[17] Parmar, Mahesh. (2020). Heart Diseases Prediction using Deep Learning Neural Network Model. International Journal of Innovative Technology and Exploring Engineering. 9. 2244-2248. 10.35940/ijitee.C9009.019320.